UNIVERSIDAD
DE LOS ANDES
VENEZUELA

Proyecto de Grado

Presentado ante la ilustre Universidad de Los Andes como requisito parcial para obtener el Título de Ingeniero de Sistemas

Diseño de interacción humano-robot para el aprendizaje de las tablas de multiplicación e implementación de los modelos de reconocimiento del habla

Por

Br. Gilberto Carrillo

Tutor: Prof. Jesús Pérez

Diciembre 2021

Diseño de interacción humano-robot para el aprendizaje de las tablas de multiplicación e implementación de los modelos de reconocimiento del habla

Br. Gilberto Carrillo

Proyecto de Grado — Sistemas Computacionales, 125 páginas
Escuela de Ingeniería de Sistemas, Universidad de Los Andes, 2021

Resumen: La estrategia tradicional que usan los niños para aprender las tablas de multiplicación mediante la memorización y repetición constante de cada una de las tablas, ha demostrado ser deficiente, ya que genera efectos negativos en los niños dificultando así el correcto aprendizaje de las tablas. Debido a este problema nace la necesidad de crear nuevas estrategias que faciliten el aprendizaje de las tablas de multiplicación. Una alternativa basada en una interacción humano-robot podría ser eficiente debido a los excelentes resultados obtenidos por los robots sociales en el contexto de la educación en los últimos años. La interacción humano-robot inherentemente requiere de alguna forma de comunicación; por lo general, esta comunicación se logra mediante dispositivos como tabletas o teclados. No obstante, si se desea que el campo de la robótica social continúe progresando hacia entornos del mundo real, debe ser incluida la interacción verbal dada la predominancia de este canal de comunicación en la interacción natural de los humanos. El proyecto de grado incluye el diseño de una interacción humano-robot para facilitar el aprendizaje de las tablas de multiplicación, así como también la implementación de los modelos de reconocimiento del habla infantil que permitirán reconocer los números y palabras que se requieran en la interacción diseñada. Además, se construirá un corpus de audio infantil para el entrenamiento y la evaluación de los modelos de reconocimiento del habla desarrollados. Este proyecto de grado es parte de un estudio más grande que busca explorar los efectos de un robot social en el aprendizaje de las tablas de multiplicación con niños.

Palabras clave: reconocimiento del habla infantil, interacción humano-robot, robótica social, tablas de multiplicación.

Índice

Índice de Tablas

Índice de Figuras

Introducción

La multiplicación es uno de los temas más importantes de las matemáticas que se enseña a los niños durante sus estudios de primaria, la cual comienza con el aprendizaje de las tablas de multiplicación. Dada la importancia de la multiplicación en la educación de los niños, se han hecho esfuerzos por crear estrategias que faciliten el aprendizaje de las tablas de multiplicación y diviertan a los niños, con la intención de que se sientan motivados durante el aprendizaje.

En los últimos años, investigaciones en el ámbito de la robótica social han demostrado los beneficios que pueden obtenerse al incorporar los robots sociales como plataformas educativas en las aulas de clases. Específicamente en el área de las matemáticas, los robots sociales logran excelentes resultados mejorando el rendimiento y estimulando la motivación de los niños durante el aprendizaje de algún tema matemático. Esto convierte a los robots sociales en una alternativa que puede servir como estrategia para ayudar y motivar a los niños durante el aprendizaje de las tablas de multiplicación.

Cuando se trata de aplicaciones donde los robots sociales interactúan con un niño, la mayoría de los investigadores se inclinan por utilizar tabletas o teclados como interfaces de comunicación con el robot. Si bien este tipo de interfaces funcionan, la comunicación no debería limitarse únicamente al uso de estos dispositivos. Agregar la capacidad de interactuar a través de la voz en este tipo de aplicaciones contribuye a interacciones más naturales con una mejor experiencia de usuario debido a que el habla es la forma más natural y eficiente que los humanos utilizan para comunicarse. Por lo tanto, el reconocimiento del habla juega un papel importante en la robótica social, ya que permite ofrecer una forma de comunicación con los robots mucho más natural e intuitiva, similar a la existente entre los humanos.

Debido a los beneficios que presentan los robots sociales en la educación, nace el interés de diseñar una interacción humano-robot que permita facilitar a los niños el aprendizaje de las tablas de multiplicación. La interacción diseñada brindará una alternativa que cubrirá las desventajas más importantes, considerando las ventajas, de las estrategias para el aprendizaje de las tablas de multiplicación aplicadas actualmente, así como también varios elementos presentes en las investigaciones de robótica social en el contexto de la educación matemática que se muestran en los antecedentes de esta investigación. Además, dada la importancia del reconocimiento del habla en este tipo de interacciones, en este trabajo de grado se considera una interacción verbal entre un niño y el robot.

En este proyecto de grado se realizará el diseño de una interacción humano-robot para el aprendizaje de las tablas de multiplicación. Del mismo modo, se hará el diseño e implementación de los modelos de reconocimiento de habla infantil según el vocabulario requerido en nuestro caso de estudio, junto con la construcción de un corpus de audio infantil en español que permitirá el entrenamiento y la evaluación de los modelos de reconocimiento del habla desarrollados.

Este proyecto de grado está organizado de la siguiente manera: en el primer capítulo se encuentra una contextualización que involucra los antecedentes de esta investigación; en el segundo capítulo se ubica el marco teórico, el cuál comprende una revisión de los conceptos necesarios para la comprensión de este trabajo; en el tercer capítulo se detalla el diseño de la interacción humano-robot para el aprendizaje de las tablas de multiplicación, incluyendo el diseño del corpus de audio infantil y el diseño e implementación de los modelos de reconocimiento del habla para el caso de estudio; en el cuarto capítulo se encuentran las pruebas que se realizaron sobre los modelos de reconocimiento del habla para el caso de estudio; finalmente, en el quinto capítulo se presentan las conclusiones y recomendaciones de este proyecto de grado.

Capítulo 1

Contextualización

El hecho de que los niños posean un dominio y fluidez de las tablas de multiplicación proporciona una gran ventaja, ya que liberan recursos mentales al momento de resolver problemas matemáticos [1], permitiendo a los niños enfocarse en el problema y en cómo deben resolverlo. Además, esto ayuda a una comprensión de conceptos más avanzados de las matemáticas durante los años de estudio en la primaria y secundaria [2].

La estrategia tradicional aplicada para que los niños aprendan las tablas de multiplicación, consiste en la memorización y repetición constante de cada una de las tablas [3, 4]. Esta estrategia se torna, en muchas ocasiones, difícil y aburrida para los niños, lo que hace que éstos se desmotiven durante el aprendizaje, entorpeciendo así el correcto aprendizaje de las tablas. Aunque en los últimos años muchas otras estrategias han sido creadas con la intención de motivar a los niños durante el aprendizaje, y han logrado obtener buenos resultados, éstas presentan una serie de desventajas, lo que conlleva a la necesidad de crear nuevas alternativas que puedan cubrir esas desventajas y logren ofrecer buenos resultados al facilitar a los niños el aprendizaje de las tablas de multiplicación.

Una alternativa que puede cubrir las desventajas más importantes presentes en las estrategias actuales para el aprendizaje de las tablas de multiplicación, considerando también las ventajas, es la robótica social. Recientemente, investigadores han examinado los efectos de robots sociales en el contexto de la educación, demostrando los efectos positivos cuando un robot social asume los roles de tutor, compañero y

aprendiz [5]. Siendo ideal que los robots se comuniquen por voz.

Un gran problema de las aplicaciones que utilizan interfaces de voz es que están desarrolladas principalmente para trabajar con el habla de los adultos. Ésto es debido a que en la actualidad la mayoría de los corpus de audio destinados al entrenamiento y evaluación de los modelos de reconocimiento del habla, se centra principalmente en el habla de personas adultas, siendo los corpus de audio infantil significativamente menos comunes [6], lo que conlleva a una carencia de datos para el entrenamiento y evaluación de modelos de reconocimiento del habla infantil. Para desarrollar aplicaciones que hagan uso del reconocimiento del habla y puedan ser utilizadas por niños de manera exitosa, es necesario poseer una gran cantidad de audios con discursos de niños para entrenar y evaluar los modelos de reconocimiento del habla basados en arquitecturas de aprendizaje profundo [7]. La mayoría de los corpus de audio infantil existentes se encuentran en el idioma inglés, mientras que para otros idiomas y dialectos, en especial los menos hablados, los corpus de audio infantil son poco comunes [8], siendo éste el caso del español venezolano.

En la siguiente sección se presentan los antecedentes de esta investigación, los cuales están conformados por investigaciones sobre estrategias para el aprendizaje de las tablas de multiplicación, investigaciones de robots sociales en el contexto de la educación aplicados al área de las matemáticas, investigaciones sobre la construcción de corpus de audio infantil e investigaciones sobre modelos de reconocimiento del habla infantil. La revisión de estos antecedentes tiene el objetivo de estudiar las diferentes estrategias utilizadas para que los niños aprendan las tablas de multiplicación, para así determinar cuáles son las ventajas y desventajas que éstas poseen y que puedan ser considerados en la interacción humano-robot a diseñar. Esta revisión también permitirá mostrar los beneficios presentes al utilizar robots sociales en el aprendizaje de las matemáticas y cómo éstos son aplicados, para así considerar los elementos más importantes al momento de diseñar la interacción. Igualmente, permitirá estudiar los métodos para la creación de corpus de audio infantil y los distintos modelos utilizados para la construcción de los reconocedores de habla infantil.

1.1 Antecedentes

Los antecedentes de esta investigación se dividen en 3 categorías: primero, investigaciones sobre estrategias para el aprendizaje de las tablas de multiplicación; segundo, investigaciones sobre robots sociales en el contexto de la educación aplicados al área de las matemáticas; y por último, investigaciones sobre la creación de corpus de audio infantil y modelos de reconocimiento del habla infantil.

1.1.1 Estrategias para el aprendizaje de las tablas de multiplicación

Debido a que la estrategia tradicional no es eficiente y por lo general no produce buenos resultados porque los niños las olvidan al momento de aprender otras operaciones matemáticas como la división [9], se han creado nuevas estrategias con la intención de lograr que los niños se motiven y se sientan atraídos en el aprendizaje de las tablas de multiplicación de una manera mucho más fácil y divertida.

Una de las estrategias usadas para motivar a los niños durante el aprendizaje de las tablas es el uso de juegos de mesa modificados para tal fin, los cuales, ayudan a mejorar en los niños la motivación, el interés y la atención por aprender. Además, permiten a los niños divertirse durante el aprendizaje, lo que hace el aprendizaje de las tablas mucho más agradable para ellos. Dentro de este tipo de estrategia podemos encontrar varios juegos de mesa como los propuestos en [10] donde se presentan los siguientes juegos: "Rompecabezas multiplicativo", que es una modificación del rompecabezas donde se arman piezas con las operaciones presentes en las tablas de multiplicación; las "Cápsulas multiplicativas", similar al juego de la memoria donde se usan chapas con las operaciones de las tablas; y "Dominó multiplicativo", una modificación del popular juego del dominó donde las fichas contienen los resultados y operaciones de las tablas. En [11] proponen una modificación del "Juego de la OCA", el cual consta de un tablero con varias casillas que contienen una operación de multiplicación. En [12] recopilan varios juegos destinados a enseñar las tablas de multiplicación, entre los cuales se encuentra el "Bingo de las tablas" basado en el bingo tradicional; en esta variación del juego se dictan operaciones de las tablas de multiplicación y las tarjetas

se llenan según el resultado de la operación dictada. La principal desventaja de este tipo de juegos es que están diseñados para permitir practicar un conjunto pequeño de las tablas de multiplicar.

Algunas estrategias plantean el aprendizaje de las tablas de multiplicación por medio de dinámicas grupales en las que los niños se divierten jugando mientras aprenden. Como en la mayoría de juegos donde los niños interactúan con compañeros de juego, presentan la ventaja de permitir a los niños desarrollar habilidades sociales como la cooperación, liderazgo, comunicación, la expresión emocional y creatividad. En [13], plantean varias estrategias para el aprendizaje de las tablas de multiplicación en forma de dinámicas con los niños entre las cuales se encuentran "Capitán multipli", la cual consiste en dar órdenes a los niños para que se formen en filas y columnas en base a una operación de la tablas de multiplicación; "Sigan la pista", en donde los niños van en orden proponiendo una operación de la tablas de multiplicación con el resultado, de manera que el siguiente debe proponer otra multiplicación en base al resultado de la anterior; "Don Pepe el pescador", que consiste en colocar sobre la pizarra varias hojas con dibujos de peces que en su reverso tienen escrita una operación de las tablas de multiplicación donde participan grupos de niños que escogen un representante por cada equipo, para que seleccione un pez del pizarrón y muestre la operación que éste contiene en su reverso a sus compañeros para que digan la operación y puedan ganar el pez. En [14], muestran una serie de actividades propuestas en una unidad didáctica para el aprendizaje de las tablas de multiplicación. Una de esas actividades es el juego denominado "Llena la cesta", que consiste en formar equipos de 4 niños entregándoles diez pelotas a cada miembro del equipo, para posteriormente encestarlas en una canasta que muestra alguna operación de la tabla del 4. Dentro de las desventajas presentes en estas estrategias, se destaca que no permiten practicar todas las tablas de multiplicación, sino sólo un conjunto pequeño de éstas.

Gracias al crecimiento que ha tenido la tecnología en los últimos años, muchas estrategias optan por hacer uso de la tecnología para cautivar con mayor facilidad a los niños, aumentando así el interés por aprender de éstos. En [15], presentan el software educativo "Tablas de multiplicar" donde los niños pueden practicar y aprender las tablas de multiplicación por medio de canciones y juegos que consisten principalmente

en la construcción de las tablas de multiplicación. En [16], implementan un software denominado "Jugando y cantando voy multiplicando" donde los niños pueden aprender las tablas de multiplicación por medio de canciones y juegos de completado de las tablas. En [17], se presenta una herramienta educacional para ayudar a los maestros a enseñar las tablas de multiplicación. La herramienta denominada "Multiplication Mat" consiste en una consola que muestra operaciones de las tablas de multiplicación, junto con una alfombrilla que captura las respuestas del niño por medio de saltos que da sobre los números dibujados en la alfombrilla. El principal problema que presentan este tipo de estrategias es que no dan una retroalimentación con la respuesta correcta cuando el niño se equivoca y optan por simplemente marcar las respuestas como incorrectas.

1.1.2 Robots sociales en la educación matemática

El uso de robots sociales en el ámbito de la educación matemática ha logrado excelentes resultados mejorando el rendimiento, aumentando la comprensión de temas matemáticos y motivando a los estudiantes durante el proceso de aprendizaje. Una técnica muy utilizada para motivar a los niños cuando interactúan con un robot es el uso de recompensas verbales, gestos y sonidos que ayuden a que los niños se sientan animados durante la interacción.

En [18], el robot social NAO actúa como asistente de enseñanza en un salón de clases con estudiantes de primaria, para apoyar el aprendizaje de una materia no robótica (aritmética) a través de una actividad innovadora basada en juegos. Aquí, cuando el robot recibe una respuesta correcta durante la actividad, recompensa al niño con gestos y sonidos, mientras que cuando recibe una respuesta incorrecta, consuela al niño y lo alienta. Los resultados de este estudio son prometedores, pues los estudiantes se mostraron encantados con el robot y, en consecuencia, indicaron un mayor interés y comprensión de los conceptos matemáticos enseñados por el robot.

En [19], presentan al robot tutor social interactivo "Ms. An" basado en la plataforma robótica NAO, el cual fue usado en sesiones de tutoría con niños para practicar problemas de multiplicación mediante la resolución de problemas y preguntas de selección múltiple, las cuales se responden a través de una interfaz basada en una tableta. Además, dependiendo del rendimiento del niño, el robot respondía con

recompensas verbales para alentarlo. Los resultados de esta investigación demostraron que los niños preferían interactuar con el robot tutor sobre otros tipos de apoyo al estudio como compañeros, programas de computadora, maestros u otros adultos.

Los robot sociales son plataformas robóticas que cautivan con facilidad a los niños. Una ventaja clave de los robot sociales es que pueden ser programados para que se adapten durante el desarrollo de las actividades dependiendo de la persona con la que interactúan y así poder brindar una experiencia personalizada.

En [20], presentan al robot NAO como un tutor inteligente adaptativo con comportamiento social. El robot fue utilizado en sesiones de tutoría con niños donde los ayudaba a resolver problemas de fracciones matemáticas por medio de pistas, que eran solicitadas por el niño a través de una tableta donde resolvía los problemas de fracciones y podía interactuar con el robot. El robot NAO proporcionaba ayuda al niño o la negaba si éste cumplía ciertas condiciones, así evitaba mal uso de las peticiones de ayuda por parte del niño. El estudio demostró que los niños que interactuaron con el robot tutor adaptativo mejoraron su rendimiento en la comprensión de las fracciones matemáticas.

En [21], se presenta un sistema de tutoría robótico autónomo que hace uso del robot NAO como tutor. El robot da lecciones cortas referentes al orden de las operaciones matemáticas mientras el niño resuelve problemas matemáticos a través de una tableta. Dependiendo del rendimiento que obtiene el niño durante la resolución de problemas el robot provee al niño tareas de descanso. Dentro de las actividades de descanso se encuentran el juego de tic-tac-toc, ejercicios físicos, ejercicios de relajación y ejercicios de reenfoque. El estudio demostró que los niños mejoraron la eficiencia y la precisión al completar problemas de matemáticas después de los descansos personalizados durante la tutoría con el robot.

En [22], presentan el diseño y la evaluación de un robot social como tutor que puede proporcionar retroalimentación de errores específicos sobre las respuestas a problemas matemáticos. En la investigación se enfocan en el dominio de la suma y la resta, para sumas y restas hasta el número 100. Se utilizó el robot NAO como plataforma robótica para llevar a cabo las sesiones de tutoría. Durante las sesiones de tutoría los niños debían interactuar con el robot y responder de manera verbal usando el

idioma neerlandés, a los problemas matemáticos planteados . Cuando los niños se equivocaban el robot decía al niño si la respuesta era correcta o incorrecta, también recibía retroalimentación detallada sobre el tipo de error que se cometió, si el robot podía identificarlo. Aunque los resultados obtenidos en la investigación no evidenciaron los efectos de aprendizaje del robot sobre el desempeño de los niños en la resolución de problemas de suma y resta, éstos expresaron haber mejorado sus habilidades gracias al robot, además de disfrutar más trabajar con el robot que cuando trabajaban con aplicaciones en tabletas.

Muy recientemente, investigadores se han centrado específicamente en estudiar los efectos de los robots sociales en el aprendizaje de las tablas de multiplicación, obteniendo resultados prometedores al mejorar el rendimiento de los alumnos. En la investigación [23], presentan al robot NAO como un robot de tutoría autónoma, donde los niños podían interactuar de manera verbal usando el idioma neerlandés con el robot, haciendo uso en parte de la técnica "Mago de Oz", mientras practican las tablas de multiplicar. Los resultados mostraron que, en promedio, los alumnos mejoraron significativamente su rendimiento incluso después de 3 tutorías de 5 minutos con el robot. Los alumnos por encima del promedio se beneficiaron más de un robot tutor con comportamiento social, mientras que aquellos por debajo del promedio se beneficiaron más de un robot que mostró un comportamiento neutral en lugar de más social.

1.1.3 Reconocimiento del habla infantil

Una razón por la que el reconocimiento del habla infantil plantea un gran desafío, es debido a la carencia de corpus de audio infantil disponibles para entrenar y probar los modelos de reconocimiento [24]. Por lo general, los corpus de audio enfocados al entrenamiento y evaluación de modelos de reconocimiento del habla se centran en el habla de personas adultas, convirtiendo a los corpus de audio infantil en un recurso escaso.

Reconocer el habla infantil presenta mayores desafíos que con el habla adulta [25], debido a que el habla de los niños posee características muy diferentes al habla de los adultos, las cuales, son atribuidas principalmente a diferencias anatómicas y morfológicas en la geometría del tracto vocal [26] y a diferencias en las habilidades

lingüísticas [24]. A nivel espectral, dado que los niños poseen tractos vocales y cuerdas vocales más pequeñas en comparación con los adultos, producen formantes y frecuencias fundamentales más altas [27]. Además, los tractos vocales de los niños cambian rápidamente a medida que maduran, lo que conlleva a una alta variabilidad en las características inter-locutor, ya que las propiedades acústicas del habla varían mucho más entre los niños que entre los adultos [28]. A nivel lingüístico, los niños tienden a reemplazar un fonema por otro y son más propensos a usar palabras imaginarias, frases gramaticalmente incorrectas y pronunciar incorrectamente las palabras [24, 29], esto a causa de que la producción del habla es una actividad motora compleja que los niños todavía están aprendiendo a dominar, por lo que la variación en la producción del habla de un mismo hablante, es decir características intra-locutor, es mucho más alta en los niños que en los adultos [28]. Es por estas razones que el rendimiento de un sistema reconocedor del habla desarrollado para adultos, disminuye drásticamente cuando se emplea para reconocer el habla infantil [30]. Igualmente, un reconocedor desarrollado para niños disminuye su rendimiento significativamente con la edad del grupo de niños usados para el entrenamiento del modelo [31].

1.1.3.1 Corpus de audio infantil

Construir un corpus de habla infantil presenta mayores retos que los corpus de habla adulta. Un desafío muy común es mantener a los niños concentrados y atentos durante el proceso de grabación, para evitar que se distraigan o se aburran durante sesiones de grabaciones muy prolongadas. A continuación, se describen varios corpus de audio infantil.

El corpus de audio TBALL [32], consiste en más de 30000 grabaciones que comprenden más de 40 horas de audio en inglés, transcritas fonéticamente utilizando el conjunto de códigos de transcripción fonética ARPABET. Las grabaciones fueron obtenidas de 256 niños divididos de manera uniforme por género, con edades comprendidas entre 5 a 8 años. La captura de las grabaciones fue realizada en escuelas utilizando un micrófono auricular para evitar el ruido ambiental, a una frecuencia de muestreo de 44100 Hz para que pudieran ser utilizadas tanto para estudios del habla de niños así como también de reconocimiento del habla. Cada sesión de grabación con

los niños tuvo una duración de 20 minutos o menos, con la intención de que los niños pudieran mantenerse concentrados el mayor tiempo posible. Uno de los principales desafíos presentes en la construcción de este corpus fue cómo motivar a los niños a decir lo que se quería que se dijera. Para esto, fue diseñado un software en Java para presentar a través de una pantalla estímulos atractivos de material legible que contenían letras, números, palabras, oraciones e imágenes, controlados por un operador humano, a fin de mantener su atención y brindarles una experiencia agradable.

NITK Kids' Speech Corpus [33] es un corpus de audio en idioma indio canarés que incluye aproximadamente 10 horas de grabaciones de habla espontánea de 160 niños, con edades comprendidas entre los 2.5 y 6.5 años. Los niños fueron divididos en 4 grupos con un intervalo de 1 año de edad, y cada grupo estaba equilibrado por género con una cantidad de 20 niñas y 20 niños por grupo. Las grabaciones fueron realizadas en una sala silenciosa utilizando un micrófono común, a una frecuencia de muestreo de 48000 Hz y 16 bits de resolución. El protocolo de captura de datos consistió en sentar a los niños en una posición cómoda frente a una computadora y un micrófono, posteriormente se les pedía describir una imagen que representaba una palabra, la cual era mostrada a éstos en una diapositiva de PowerPoint en la pantalla de la computadora. Como los niños se aburren con facilidad, se dieron descansos suficientes entre las grabaciones para mantener respuestas adecuadas durante la sesión de grabación.

CHOREC [34], es un corpus de audio de discurso leído y etiquetado fonéticamente, obtenido de 400 niños de primaria hablantes del idioma neerlandés con edades entre los 6 y 12 años. De éstos, 274 niños pertenecían a escuelas regulares y el resto pertenecía a escuelas para niños con discapacidades de aprendizajes específicas. Las grabaciones fueron realizadas en una habitación silenciosa con la intención de evitar el ruido. El discurso de los niños se grabó a 22050 Hz por medio de 2 micrófonos: un micrófono auricular y un micrófono de escritorio. Para las grabaciones, se mostraban a través de un computador listas de palabras que se presentaban de forma individual e historias que se presentaban párrafo por párrafo. Los niños fueron instruidos para que intentaran leer las palabras y párrafos presentados en la pantalla, con la mayor precisión y fluidez posible. Se realizaron 3 sesiones de grabación por niño, cada sesión tenía un máximo de

20 minutos de duración. La mayoría de los niños estaban muy motivados y ansiosos por participar en las grabaciones. Sin embargo, algunos niños, que a menudo tienen serias dificultades de lectura, debían estar motivados por la promesa de que luego podrían escuchar sus propias grabaciones.

En [8], se describe el corpus de audio CNG que consta de 21 horas de discurso leído en portugués, grabado de 510 niños con edades comprendidas entre 3 a 10 años de edad. Las grabaciones consisten en un total de 292 frases ricas fonéticamente, notas musicales, números y secuencias de números que fueron breves y no incluían palabras difíciles. Para la captura de las grabaciones se utilizó una interfaz web donde se presentaba a los niños el texto a grabar. Las sesiones de grabación fueron supervisadas por seis personas capacitadas para la tarea, y tuvieron lugar en una habitación tranquila. En el caso de los niños que tuvieron problemas para leer el texto, el supervisor de grabación las leyó primero y luego los niños las repitieron. Cada grabación fue captada usando un micrófono auricular con cancelación de ruido a una frecuencia de muestreo de 22000 Hz y 16 bits de resolución.

En [35], se construye el corpus de audio CID children's speech corpus que fue recolectado de 436 niños equilibrados por género, con edades comprendidas entre 5 y 18 años, y 56 adultos. Las grabaciones que consistían en 15 palabras y 5 frases fonéticamente ricas, se realizaron en una cabina con tratamiento de sonido ubicada dentro de una caja de panel de vidrio, utilizando un micrófono de alta fidelidad a una frecuencia de muestreo de 20000 Hz y 16 bits de resolución. Las palabras y frases se presentaron en un monitor de computadora dos veces en orden aleatorio mientras se le pedía a los participantes que repitieran el texto presentado. Aquellos niños que presentaban problemas para leer las palabras y frases, realizaban las grabaciones mediante la imitación de una muestra pregrabada por una patóloga del habla femenina. El corpus de audio final consta de 24152 archivos de audios en formato WAV transcritos fonéticamente.

OGI Kids' Speech [36] es un corpus de audio que consiste en una colección de grabaciones de discursos espontáneos y lectura en inglés. Las grabaciones están conformadas por 205 palabras aisladas, 100 oraciones solicitadas y 10 cadenas numéricas, obtenidas de 1100 niños con edades comprendidas entre los 5 y 15 años. Un

grupo de 100 niños por grado divididos equilibradamente por género, pertenecientes a los grados comprendidos entre preescolar y el décimo grado, participó en la recolección de las grabaciones. La captura de las grabaciones fue realizada durante las horas de clases de los estudiantes por lo que las sesiones de grabación tenían una duración máxima de 30 minutos. El protocolo de grabación consistía en mostrar en una pantalla de computador un texto y reproducir una muestra pregrabada con voz humana del texto, en sincronía con la animación facial usando el personaje animado en 3D "Baldi", posteriormente, el estudiante repetía el texto. Una vez completada esta fase, el experimentador realizaba al estudiante una serie de preguntas destinadas a provocar un discurso espontáneo, los cuales fueron transcritos ortográficamente según las convenciones de transcripción CSLU. La cantidad total de discurso grabado por estudiante fue aproximadamente de 8 a 10 minutos. Durante las grabaciones se utilizó un micrófono a una frecuencia de muestreo de 16000 Hz y 16 bits de resolución.

1.1.3.2 Modelos para el reconocimiento del habla infantil

Una forma de desarrollar un reconocedor del habla infantil es utilizando alguno de los frameworks de software libre disponibles en Internet, como HTK [37], CSLU [38] o Kaldi [39]. Estos frameworks utilizan métodos estadísticos basados en modelos ocultos de Markov (HMM) para adaptarse a la variabilidad en los patrones de voz.

En [36], desarrollan un sistema reconocedor de habla infantil utilizando el framework CSLU Toolkit y el corpus de audio infantil OGI Kids' Speech. Para el entrenamiento de este reconocedor se usaron las palabras aisladas de los grados 2 a 10 del corpus de audio, junto con ejemplos obtenidos del corpus de audio TIMIT [40] para contrarrestar la falta de datos. Para evaluar el desempeño del reconocedor se realizaron dos tipos de pruebas. La primera usando un conjunto de prueba de 205 palabras aisladas en la que se obtuvo una precisión de 97.5%. La segunda usando un conjunto de prueba de 100 palabras que no fueron usadas para el entrenamiento, con la cual se obtuvo una precisión del 37.9%.

En [41], se presenta un modelo reconocedor de habla infantil orientado a niños malayos, el cual fue desarrollado con el framework HTK toolkit. El modelo fue entrenado usando 360 frases de un pequeño corpus de audio que comprende las voces

de seis niños que pronuncian un total de 390 frases conformadas por 1404 palabras, de las cuales, 987 son palabras que no se repiten. En los resultados de la prueba de rendimiento, el modelo muestra una precisión a nivel de oración del 71.51%. A nivel de palabras, la precisión es del 76.70%, que se basa en un total de 160 palabras disponibles en las 30 frases que conformaban el conjunto de prueba. En total, el modelo obtuvo una tasa de error de palabra del 23.30%. El modelo desarrollado en esta investigación puede reconocer el habla infantil de niños hablantes del idioma malayo a un nivel aceptable a pesar del uso de un pequeño corpus de audio.

El problema de utilizar un framework basado en modelos ocultos de Markov para desarrollar un reconocedor del habla infantil es que estos conjuntos de herramientas carecen de los algoritmos y técnicas de alto rendimiento que se han desarrollado en los últimos años. Actualmente, los mejores resultados en un reconocedor del habla infantil se obtienen cuando se desarrollan utilizando técnicas de aprendizaje profundo.

En [7], se construye un reconocedor del habla de vocabulario extenso para niños y adultos, que se utiliza en la aplicación móvil YouTube Kids. El conjunto de entrenamiento se obtuvo mediante un muestreo del tráfico de búsqueda por voz de Google, donde se capturaron 1.9 millones de declaraciones hechas por niños y 2.6 millones hechas por adultos. Además, para lograr que el reconocedor fuera robusto al ruido, se creó un conjunto de datos que contenía ruidos de fondos artificiales. Para el modelo acústico del reconocedor se compararon las redes neuronales recurrentes de memoria a corto plazo y a largo plazo (LSTM) con las redes neuronales convolucionales profundas LSTM (CLDNN). Los resultados de este trabajo demostraron que las redes neuronales CLDNN superan a una red neuronal LSTM en una variedad de condiciones diferentes, obteniendo una tasa de error de palabra del 12.5% para el habla de los adultos y del 9.4% para el habla de los niños. En los resultados para la prueba en condiciones ruidosas, el reconocedor obtiene una tasa de error de palabra del 9.4% en condiciones sin ruido y del 11.8% en condiciones ruidosas para el habla de los niños, mejorando el reconocimiento en condiciones ruidosas sin afectar el rendimiento en condiciones menos ruidosas.

Un problema presente al momento de desarrollar reconocedores del habla infantil es la falta de grandes cantidades de datos para el entrenamiento de los modelos. Varias

técnicas de aumento de datos pueden ser empleadas para contrarrestar este problema, como por ejemplo utilizar técnicas para transformar voces pertenecientes a corpus de audio de adultos, los cuales se encuentran en mayor cantidad que los de niños.

En [42], se explora el aumento de datos para el reconocimiento del habla de los niños utilizando el mapeo de características estocásticas (SFM) para transformar audios de adultos. Se realizaron experimentos utilizando el corpus de audio en inglés PF-STAR [43], WSJCAM0 [44] y ABI [45]. Con el corpus de audio PF-STAR se entrenó un modelo base utilizando Kaldi, con el que se obtuvo una tasa de error de palabra del 29%, posteriormente se aplicó SFM en las bases de datos WSJCAM0 y ABI para entrenar dos modelos más y evaluar su rendimiento. Los mejores resultados fueron obtenidos con el modelo entrenado con PF-STAR y WSJCAM0 aplicando SFM, obteniendo una tasa de error de palabra del 27.2%, una mejora relativa del 6.2% sobre el modelo base, demostrando que los reconocedores de habla infantil pueden hacer uso de datos de adultos cuando son transformados con SFM.

En [46], utilizan la normalización de la longitud del tracto vocal (VTLN) para reducir el desajuste entre los datos del corpus de audio TIMIT y el corpus de audio infantil CMU Kids. En la investigación se observó que el habla de mujeres adultas se parece mucho más a la de los niños. Por lo tanto, se seleccionaron sólo datos de mujeres adultas de TIMIT para aumentar el conjunto de entrenamiento usando VTLN. Al entrenar un modelo con el conjunto de entrenamiento aumentado, se observó que la tasa de error de palabra era de 16.92% comparado con un 19.50% de un modelo base entrenado solo con los datos de los niños, demostrando que VTLN es eficaz para tratar la variabilidad entre el habla de niños y adultos.

Aunque los corpus de audio infantil son menores que los corpus de audio adulto, varios corpus de audio infantil pueden ser usados para aumentar los datos. En [47], se desarrollan varios modelos utilizando el framework Kaldi Toolkit para reconocer el habla infantil de niños jamaicanos. El objetivo de la investigación fue explorar el aumento de datos utilizando corpus de audio infantil de dialectos del inglés relacionados al inglés jamaicano. Se utilizaron 3 corpus de audio infantil para el estudio, CMU Kids, PF-STAR y el corpus de audio de inglés jamaicano JAMLIT para entrenar 3 modelos bases. Para el modelo base entrenado con la base de datos JAMLIT se obtuvo una

tasa de error de palabra del 22.9%, con el modelo entrenado con PF-STAR se obtuvo una tasa de error de palabra del 85.7%, finalmente, con el modelo entrenado con CMU Kids se obtuvo una tasa de error de palabra de 84.1%. Posteriormente se aplicó el aumento de datos con el corpus de audio en inglés británico PF-STAR y americano CMU Kids. Cuando al modelo entrenado con el corpus de audio PF-START se le agregó una fracción del corpus de audio JAMLIT se obtuvo una tasa de error de palabra del 25.6%, obteniendo una mejora del 58.1% en comparación con un modelo entrenado solo con el corpus de audio PF-STAR. Con CMU Kids, la mejora obtenida fue del 59.6%, obteniendo una tasa de error de palabra del 24.9%. El estudio muestra que los modelos creados con datos de niños de un dialecto del inglés no reconocen idealmente el habla de un dialecto relacionado. Además, demuestra que se pueden utilizar datos no modificados del dialecto de destino para aumentar los datos de un dialecto relacionado para producir un modelo que funciona casi tan bien como un modelo completo del dialecto de destino.

Algo a tomar en consideración cuando se estudia el reconocimiento del habla infantil en escenarios del mundo real, especialmente en escenarios de interacción niño-robot de aplicaciones educativas, es que se debe lidiar con el ruido de fondo presente en las aulas de clase para que los modelos de reconocimiento funcionen correctamente en condiciones ruidosas. Una forma de lidiar con este problema es aplicar técnicas de aumento de datos utilizando audios que contengan ruidos de fondo similares a los esperados en el entorno donde se desarrollará la interacción.

En [48], implementan un modelo de reconocimiento del habla utilizando el framework Kaldi Toolkit que puede ser usado tanto por niños como por adultos en aplicaciones donde se interactúa con robots. Para el entrenamiento del modelo se usaron dos corpus de audio en inglés británico. El primero es la versión en inglés británico WSJCAMO, el segundo fue el corpus de audio infantil en inglés británico PF-STAR. Además, para mejorar la robustez del reconocedor en entornos ruidosos, utilizaron audios de ruido de fondo de la base de datos CHiME [49], en concreto, ruido de fondo presente en una cafetería para aumentar los datos de entrenamiento, ya que éste era el que más se ajustaba al entorno donde el robot se pondría a prueba. Al evaluar el reconocedor con los conjuntos de prueba, se obtuvo una tasa de error de palabra del

9.9% en condiciones silenciosas y 13.1% para condiciones ruidosas con el habla infantil, para el habla de los adultos se obtuvo un 4.9% en condiciones silenciosas y 12.3% en condiciones ruidosas. Además, para determinar el rendimiento del reconocedor del habla en tiempo real, el modelo se probó en un evento de un museo público como parte de un estudio de investigación en el que 320 niños interactuaron con un robot, logrando un 90% de precisión al reconocer frases dichas por los niños.

La mayoría de los reconocedores del habla infantil son desarrollados para utilizar un vocabulario extenso, pero en muchos escenarios de interacción, no es necesario un sistema reconocedor de un vocabulario extenso para que la interacción se lleve a cabo de manera correcta, en estos casos, el reconocimiento de palabras clave se puede aplicar como una alternativa, reconociendo comandos o algunas palabras para mantener una interacción.

En [50], desarrollan un modelo reconocedor de palabras clave presentes en el habla infantil en un escenario de interacción niño-robot, haciendo uso de redes neuronales bidireccionales de memoria a corto plazo y a largo plazo (BLSTM) para la detección de fonemas, junto con una red bayesiana dinámica. Para el entrenamiento del modelo se consideró un conjunto de datos que contenía 25 palabras clave diferentes obtenidas del corpus de audio FAU Aibo Emotion Corpus [51], un corpus de audio infantil en alemán con grabaciones de niños de 10 a 13 años que se comunican con un robot mascota. Según las pruebas realizadas en esta investigación, el reconocedor de palabras clave propuesto logró una tasa de detección de hasta el 95.9%. Además, el ser basado en fonemas permite que sea independiente del vocabulario, añadiendo la versatilidad de agregar nuevas palabras sin la necesidad de reentrenar el modelo.

Si bien los reconocedores de palabras clave basados en fonemas poseen la versatilidad de ser independientes del vocabulario, éstos requieren una gran cantidad de audios etiquetados fonéticamente. Los reconocedores de palabras clave basados en palabras solo necesitan de grabaciones que contengan las palabras clave de ejemplos. Además, éstos llegan a tener un mejor rendimiento tanto en precisión, como en menor tiempo de latencia [52].

En [53], desarrollan un reconocedor de palabras clave para ser usado en Mole madness [54], un juego controlado por voz en el que dos jugadores en configuración

niño-niño o niño-robot, mueven un topo virtual a través de su entorno, obteniendo recompensas y evitando obstáculos. En el juego, un jugador debe crear el movimiento horizontal del topo usando la palabra GO y el otro debe crear movimiento vertical con la palabra JUMP. Aunque se presenta una tarea simple de reconocimiento de dos palabras clave, este escenario conlleva a varios desafíos como ruido de fondo, habla superpuesta y variabilidad léxica de los niños. Para el entrenamiento del reconocedor de palabras clave se recopilaron grabaciones de 62 niños de entre 5 y 10 años que jugaron Mole Madness, las cuales fueron segmentadas y transcritas a nivel de palabra. En las pruebas realizadas al reconocedor de palabras clave en las mejores condiciones se obtuvo una precisión del 89% en la configuración niño-niño y del 93% en la configuración niño-robot.

1.2 Planteamiento del problema

La estrategia tradicional con la que los niños aprenden las tablas de multiplicación, por medio de la memorización y repetición constante, ha demostrado no ser efectiva ya que los niños muestran deficiencias al retener a largo plazo en su memoria las tablas de multiplicación [55], olvidándolas semanas o incluso días después. Además, esta manera de aprender las tablas de multiplicación genera una gran presión emocional al niño aprendiz debido a la insistencia en la memorización, lo que conlleva a un rechazo y actitud negativa a otros temas referentes con las matemáticas [56]. Aunado a esto, es considerada difícil y aburrida por los niños, lo que hace que éstos se desmotiven durante el aprendizaje.

Debido a las desventajas que presentan las estrategias actuales, es necesario encontrar nuevas formas de ayudar a los niños con el aprendizaje de las tablas de multiplicación que puedan mitigar esos problemas. Dado que el campo de la robótica social en el contexto de la educación presenta excelentes resultados e investigaciones recientes muestran los efectos positivos que generan los robots sociales en el área de las matemáticas [18, 19, 20, 21, 22, 23], pareciera que una interacción humano-robot sería una buena alternativa para facilitar a los niños el aprendizaje de las tablas de multiplicación.

Por lo general, en las aplicaciones donde se hace uso de robots sociales, la interacción con el robot se lleva a cabo mediante enfoques de interacciones que no se basan en la interacción verbal, siendo los más populares el uso de tabletas y teclados, los cuales funcionan muy bien para interacciones no verbales. Sin embargo, como se explica en [57], si se desea que el campo de la robótica social continúe progresando hacia entornos del mundo real, no es realista excluir la interacción verbal debido a la predominancia de este canal de comunicación en la interacción natural de los humanos, por lo tanto, es necesario desarrollar enfoques de interacciones que hagan uso de la comunicación verbal. Una forma de lograr una interacción verbal con un robot es mediante el reconocimiento del habla [58], que permite traducir señales de audio en una entrada legible por las computadoras. En este sentido, es necesario que las interacciones con robots sociales permitan la interacción verbal, ya que así se explotan aún más las habilidades de un robot social.

1.3 Justificación

La robótica social es un área que cobra cada vez más importancia en el contexto de la educación a través de diferentes aplicaciones tales como: enseñar a un niño un segundo idioma [59], enseñar a un robot a escribir a mano [60], narrar cuentos a niños de preescolar para mejorar su rendimiento cognitivo/motor [61], logrando efectos positivos en los niños. Varios estudios donde hacen uso de robots sociales para el aprendizaje de las matemáticas presentados en la sección de antecedentes, demuestran que el uso de estos sistemas robóticos motivan e involucran a los niños en el aprendizaje de temas matemáticos, logrando que los niños se diviertan y disfruten al interactuar con un robot social.

Es por esta razón, que nace el interés de brindar una nueva forma de aprendizaje de las tablas de multiplicación a través de una interacción humano-robot, de tal forma que permita cubrir las desventajas más importantes, considerando las ventajas, de las estrategias que se aplican actualmente, incluyendo también algunos elementos aplicados en el uso de robots sociales en el aprendizaje de las matemáticas. Además, en aras de que la robótica social continúe progresando hacia entornos del mundo real, se tomará en

cuenta un enfoque de comunicación verbal entre el niño y el robot durante la interacción mediante el diseño y la implementación de modelos de reconocimiento del habla infantil que permitan reconocer números y palabras.

1.4 Objetivos

1.4.1 Objetivo general

Diseñar una interacción humano-robot para el aprendizaje de las tablas de multiplicación e implementar los modelos de reconocimiento del habla.

1.4.2 Objetivos específicos

- Investigar las estrategias actuales utilizadas para el aprendizaje de las tablas de multiplicación.
- Estudiar los elementos característicos de los robots sociales en el contexto de la educación matemática.
- Diseñar la interacción humano-robot para el aprendizaje de las tablas de multiplicación.
- Construir el corpus de audio infantil para el entrenamiento y evaluación de los modelos de reconocimiento del habla.
- Implementar los modelos de reconocimiento del habla infantil según el vocabulario requerido en la interacción diseñada.
- Evaluar el rendimiento de los modelos de reconocimiento de habla infantil implementados.

1.5 Metodología

1.5.1 Fase de diagnóstico

- Revisar documentos científicos relacionados con las estrategias utilizadas para el aprendizaje de las tablas de multiplicación.
- Revisar documentos científicos relacionados con los robots sociales en el contexto de la educación matemática.
- Revisar documentos científicos relacionados con modelos para el reconocimiento del habla infantil.
- Revisar documentos científicos relacionados con la construcción de corpus de audio infantil.

1.5.2 Fase de diseño

- Diseñar la interacción humano-robot para el aprendizaje de las tablas de multiplicación.
- Diseñar el protocolo de captura de audio para el corpus de audio infantil.
- Diseñar los modelos de reconocimiento del habla infantil.

1.5.3 Fase de implementación

- Crear el corpus de audio infantil.
- Implementar los modelos de reconocimiento del habla infantil.

1.5.4 Fase de pruebas

- Realizar pruebas sobre los modelos de reconocimiento del habla infantil para evaluar la tasa de reconocimiento.

1.6 Alcance

En este proyecto de grado se realizará el diseño de una interacción humano-robot para el aprendizaje de las tablas de multiplicación, que logre cubrir las desventajas más importantes presentes en las estrategias actuales para el aprendizaje de las tablas de multiplicación, tomando en consideración las ventajas presentes en estas estrategias, así como también los elementos más importantes considerados en las investigaciones de robótica social en el contexto de la educación matemática que se muestran en los antecedentes de esta investigación.

De igual forma, el proyecto de grado incluye el diseño e implementación de los modelos de reconocimiento del habla infantil que permitirán reconocer los números y palabras requeridas en la interacción diseñada, junto con la creación de un corpus de audio infantil para el entrenamiento y evaluación de los modelos desarrollados. El corpus de audio infantil para implementar los modelos de reconocimiento del habla será creado a partir de muestras de audio provenientes de niños de primaria que se presten como voluntarios a esta investigación.

Realizar un estudio en profundidad de los efectos en el aprendizaje de los niños utilizando un robot social, es una tarea desafiante que requiere de un proceso de varias etapas, el cual puede tomar mucho tiempo. Este proyecto de grado hace parte de un estudio más grande que busca explorar los efectos de un robot social en el aprendizaje de las tablas de multiplicación con niños. Por lo tanto, en el proyecto de grado se limitará a diseñar la interacción humano-robot, así como también diseñar e implementar los modelos de reconocimiento del habla necesarios para permitir la interacción verbal durante la interacción.

Capítulo 2

Marco teórico

2.1 Robótica social

En los últimos años, ha habido un interés creciente en el desarrollo de robots que interactúan de forma autónoma con los humanos siguiendo reglas de comportamiento social, con la intención de que éstos se acoplen a la vida cotidiana de las personas. La robótica social es el estudio de los robots que interactúan y se comunican entre ellos, con los humanos y con el medio ambiente, dentro de la estructura social y cultural adjunta a sus roles [62]. Recientemente los robots sociales han sido utilizados en diferentes aplicaciones para resolver problemas complicados, donde humanos y robots interactúan de manera natural e interpersonal para obtener ventajas de su colaboración, como por ejemplo en: hogares [63], educación [5] y salud [64].

2.1.1 Robot social

El concepto de robot social continúa hasta el día de hoy bajo debate. Varios autores han dado su punto de vista y, aunque actualmente no existe un consenso en torno a una definición precisa, en general [65, 66, 67, 68] están de acuerdo en que un robot social posee las siguientes características:

- **Encarnación física**: un robot social debe tener un cuerpo físico con el cual interactuar.

- **Sociabilidad**: un robot debe ser capaz de interactuar con las personas mostrando rasgos similares a los humanos mientras sigue las reglas sociales (definidas a través de la sociedad) asociadas a su función.
- **Autonomía**: un robot social debe tomar decisiones por sí mismo.

A continuación se presentan las definiciones más relevantes encontradas en la literatura.

Bartneck y Forlizzi [65] proponen la siguiente definición: un robot social es un robot autónomo o semiautónomo que interactúa y se comunica con los humanos siguiendo las normas de comportamiento que esperan las personas con las que se pretende que interactúe el robot. Esta definición implica 3 condiciones:

1. Un robot social tiene una encarnación física. Los personajes de pantalla o cualquier tipo de agente virtual quedarían excluidos por esta definición.
2. La autonomía es un requisito para un robot social. Un robot semiautónomo puede definirse como social si comunica un conjunto aceptable de normas sociales. Un robot completamente controlado a distancia no puede considerarse social, ya que no toma decisiones por sí mismo y es simplemente una extensión de otro humano.
3. La comunicación y la interacción con los humanos es un punto crítico en esta definición. Por tanto, los robots que solo interactúan y se comunican con otros robots no se consideran robots sociales. Es probable que la interacción sea cooperativa, pero no se limita a ella.

Desde el punto de vista de Fong et al. [66], los robots sociales son agentes encarnados que forman parte de un grupo heterogéneo: una sociedad de robots o humanos. Son capaces de reconocerse y participar en interacciones sociales, poseen historias (perciben e interpretan el mundo en términos de su propia experiencia), y se comunican explícitamente y aprenden unos de otros. Por lo tanto, un robot social posee las siguientes características sociales humanas:

1. Tiene que ser capaz de expresar y percibir emociones.
2. Comunicarse con diálogos de alto nivel.
3. Aprender y reconocer modelos de otros agentes.
4. Debe ser capaz de establecer y mantener relaciones sociales, utilizando señales naturales (mirada, gestos, etc.) y exhibiendo personalidad y carácter distintivos.
5. El robot también debe desarrollar competencias sociales.

Por otro lado, para Hegel et al. [67], un robot social combina aspectos técnicos y sociales, pero los aspectos sociales son el propósito central de los robots sociales. El robot no es un robot social per se, sino que necesita capacidades comunicativas específicas para convertirse en un robot social. En primer lugar, implica que el robot se comporte (funcione) socialmente dentro de un contexto y, en segundo lugar, implica que el robot tenga una apariencia (forma) que exprese explícitamente ser social en un aspecto específico para cualquier usuario. Desde este punto de vista, un robot social contiene un robot y una interfaz social. Una interfaz social incluye todas las características diseñadas por las cuales un usuario juzga que el robot tiene cualidades sociales. Esta interfaz está constituida por los siguientes componentes:

- **Forma social**: elementos que contribuyen a la comunicación humano-robot, como el rostro.
- **Función social**: todos los aspectos que computan cualquier comportamiento social artificial de un robot social son parte de las funciones sociales. Por ejemplo, las emociones artificiales, los módulos para el reconocimiento y la producción de voz son funciones que producen y alteran la interacción social.
- **Contexto social**: una aplicación es un contexto de un robot e influye en la función. Dentro de una aplicación, un robot debería poder realizar todas las tareas necesarias para mantener las expectativas de un usuario. Pero el robot no tiene que ser capaz de hacer cosas fuera de su aplicación prevista, porque en general la gente no espera que el robot haga cosas para las que no

está preparado. Por tanto, las aplicaciones son un criterio para disminuir la complejidad y determinar las tareas de los robots sociales.

Para Breazeal [68] un robot social es capaz de comunicarse con los humanos, entenderlos e incluso relacionarse con ellos, de manera personal. Debería poder entender a los humanos y a sí mismo en términos sociales. A su vez, los seres humanos deberían poder entender al robot en los mismos términos sociales: poder relacionarse con el robot y empatizar con él. Dicho robot debe ser capaz de adaptarse y aprender a lo largo de su vida, incorporando experiencias compartidas con otras personas en su comprensión de sí mismo, de los demás y de las relaciones que comparten. En resumen, un robot sociable es socialmente inteligente de una manera similar a la humana. Para alcanzar este objetivo de un robot social, Breazeal establece los siguientes requisitos:

1. Un robot social debe estar encarnado de manera situada, ya que la experiencia social depende de entornos simétricos donde las entidades interactúan entre sí.

2. Un robot social debe tener cualidades reales, ya que los humanos tienden a antropomorfizar la tecnología y a interpretar el comportamiento como intencional.

3. Un robot social debe ser capaz de identificar quién es la persona, con quién está interactuando, qué está haciendo y cómo lo está haciendo.

4. Un robot social debe ser entendido, esto significa que el ser humano necesita ser capaz de leer las actividades (expresiones, mímica, etc.) del robot.

5. Un robot social, al igual que los humanos, también debe poder aprender continuamente sobre sí mismo, aquellos con quienes interactúa y su entorno. De esta forma las nuevas experiencias darán forma continuamente a la historia personal del robot e influirán en su relación con los demás.

2.1.2 Interacción humano-robot

Tradicionalmente, la interacción entre humanos y robots se llevaba a cabo principalmente de forma unidireccional, es decir, controles simples de encendido y apagado o joysticks analógicos para operar articulaciones de manipuladores y vehículos

remotos. Con el tiempo, a medida que los robots se han vuelto más inteligentes, la naturaleza de la comunicación entre humanos y robots se ha vuelto cada vez más parecida a la relación entre dos seres humanos y menos parecida a la del uso de una herramienta.

La interacción humano-robot (HRI por sus siglas en inglés) se puede definir como el estudio de los humanos, los robots y las formas en que se influyen entre sí. Como disciplina, HRI considera el análisis, diseño, modelado, implementación y evaluación de robots para uso humano [69]. En este proyecto de grado es de interés la interacción humano-robot debido a que la nueva estrategia para el aprendizaje de las tablas de multiplicación esta destinada a ser implementada en un escenario donde interactúan de manera social un niño y un robot.

2.1.2.1 Modelo de interacción humano-robot MIHR

El desarrollo de aplicaciones basadas en robots sociales no solo se enfrenta a los desafíos convencionales de la robótica, como la localización del robot o la planificación del movimiento. Un desafío importante en el desarrollo de aplicaciones que hacen uso de robots sociales es la organización y la forma en que se comunican cada uno de los componentes de software que facilitarán el desarrollo de las habilidades sociales del robot al momento de interactuar con los humanos.

A continuación se presenta el modelo de interacción humano-robot MIHR (ver Fig. 2.1), propuesto en [70] que fue tomado como base para la gestión de la dinámica de interacción del robot social en nuestra estrategia para el aprendizaje de las tablas de multiplicación.

El modelo MIHR, es un modelo que describe la interacción entre un humano y un robot, el cual toma como base un modelo de interacción humano-humano [71] para lograr una interacción parecida a la interacción existente entre personas. MIHR está conformado por tres elementos principales: la modalidad de la comunicación humana, la adaptación de la interacción y la expresión de emociones.

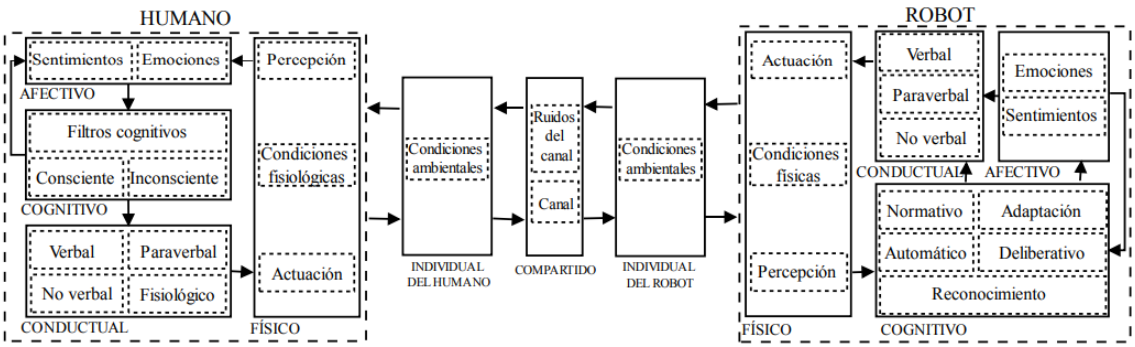


Figura 2.1: *Modelo de interacción humano-robot MIHR [70]*

- **Modalidad de la comunicación humana**: es la forma en la que los robots sociales se comunican y dado que la comunicación en lenguaje natural usa diferentes formas para comunicar una misma intención, estos pueden hacerlo de diferentes formas, como por ejemplo: la voz, la mirada, gestos o el contacto físico.

- **Capacidad de adaptación**: los robots sociales deben ser capaces de personalizar su comportamiento con la intención de proporcionar interacciones apropiadas en lugar de brindar interacciones genéricas que se mantengan constantes en todas las personas o en una misma persona, ya que las interacciones con personas requieren considerar que una misma persona puede tener un comportamiento inconsistente en el tiempo, incluso, mientras interactúa con el robot.

- **Expresión de emociones**: los robots sociales necesitan entender y expresar estados afectivos para mejorar sus interacciones, esto con la finalidad de ayudar a que el robot comunique su estado emocional, alentar los comportamientos deseados de las personas o ayudar a las personas a conectarse emocionalmente con el robot.

MIHR está constituido por dos niveles: un nivel interno, que describe el proceso interno de cada participante en la interacción desde que percibe información hasta que genera respuestas, y un nivel externo, que describe cómo puede ser alterada la información desde que es emitida por un participante hasta que es recibida por otro.

2.1.2.1.1 Nivel interno del humano El nivel interno del humano está constituido por cuatro módulos: físico, afectivo, cognitivo, y conductual. El nivel interno recibe la información que proviene del nivel externo, modificándose según los sentidos de la persona y activando estados afectivos según la interpretación dada. Posteriormente, se hace una nueva interpretación según los filtros cognitivos de la persona, para determinar la intención de comunicación de la otra persona, activando estados afectivos nuevamente. Después, se genera el objetivo y el estilo de comunicación, a partir de los aportes realizados inconsciente y conscientemente, y de los estados afectivos. Por último, a la respuesta se le asignan los componentes conductuales, según el objetivo y el estilo de comunicación, restringidos por las condiciones fisiológicas de la persona.

2.1.2.1.2 Nivel externo Durante la ejecución de una interacción la persona envía información al robot y viceversa. Esa información es enviada a través de un canal de comunicación y esta puede llegar a ser alterada por el canal durante el recorrido. Además, la información puede ser alterada durante el camino que recorre desde que es generada por el emisor hasta que es incorporada al canal de comunicación o durante el proceso de recepción mientras es desincorporada del canal de comunicación.

El nivel externo se comporta de la misma forma para el humano y el robot ya que ambos pueden cumplir el rol de emisor y receptor. El nivel externo se compone de tres módulos: módulo individual del humano, módulo individual del robot y compartido. El nivel externo funciona de la siguiente manera: en primer lugar se altera la información de entrada según las condiciones ambientales presentes entre el emisor y el canal (módulo individual del emisor). Posteriormente, incorpora la información al canal de comunicación y la modifica según los ruidos del canal (módulo compartido). Por último, desincorpora la información del canal (módulo individual del receptor) y la modifica según las condiciones ambientales presentes entre el canal y el receptor, obteniendo así la salida de este nivel.

2.1.2.1.3 Nivel interno del robot La entrada de este nivel proviene del nivel externo la cual es recibida a través de los sensores del robot. Posteriormente, se aplican algoritmos de reconocimiento, con la intención de interpretar los datos y descubrir la intención de la comunicación. Dentro de este nivel se gestionan dos tipos de respuestas:

automáticas y deliberadas. Las respuestas automáticas se ejecutan inmediatamente después de descubrir que son requeridas, mientras que las respuestas deliberadas están basadas en un objetivo de comunicación, el cual es determinado según la intención, las normas de interacción y el estado afectivo, para luego adaptarlo según la persona y generar las formas verbales, paraverbales y no verbales que serán traducidas en señales que entiendan los actuadores para que sean ejecutadas. A continuación se describe con detalle cada módulo presente en el nivel interno del robot:

- **Módulo físico**: este módulo está compuesto por tres componentes: percepción, actuación y condiciones físicas. Si la información proviene del módulo individual del robot del nivel externo, el componente de percepción se encarga de capturar las señales según lo indicado por el componente de condiciones físicas (repetibilidad, sensibilidad, etc.) acerca de los sensores, de manera multimodal para detectar los datos de interés, para luego enviarlos al componente de reconocimiento del módulo cognitivo. Por otro lado, si la información que recibe este módulo proviene del módulo conductual, el componente de actuación se encarga de traducir las órdenes en señales comprendidas por los actuadores, para posteriormente enviarlas a que sean ejecutadas según lo indicado acerca de los actuadores por el componente de condiciones físicas (precisión, velocidad, etc.).

- **Módulo cognitivo**: está compuesto por cinco componentes: reconocimiento, automático, deliberativo, normativo y adaptación. En el componente de reconocimiento se aplican algoritmos de reconocimiento (de personas, voz, temas de conversación, lugares, estados afectivos, eventos, entre otros), usando los datos obtenidos desde el módulo físico. En el componente automático se identifican los eventos que requieren respuesta inmediata y se envían las respuestas directamente al módulo conductual, con el propósito de que sean ejecutadas rápidamente. En el componente deliberativo se aplican algoritmos que permiten interpretar lo que se ha reconocido en los diferentes modos o canales, para identificar la intención de comunicación y generar un objetivo de comunicación según las normas de interacción y el estado afectivo. El componente normativo provee las normas sociales de la interacción. En el componente de adaptación, se adecúa la respuesta

según el objetivo y la persona. Adicionalmente, el módulo cognitivo envía los resultados de la intención de comunicación identificada al módulo afectivo.

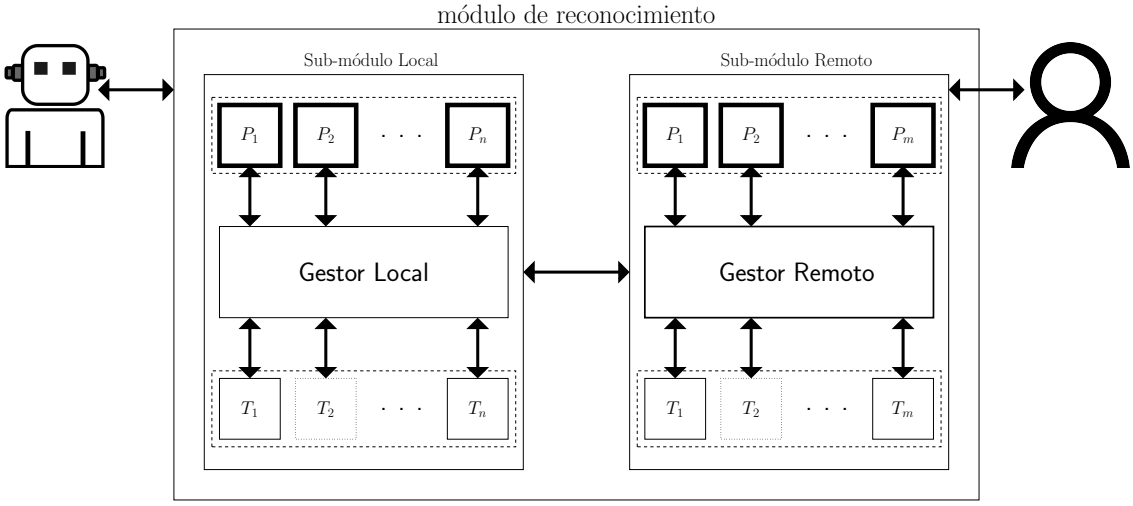


Figura 2.2: *Módulo de reconocimiento*

– **Módulo de reconocimiento**: el módulo de reconocimiento [72] es un subcomponente del módulo cognitivo que se encarga de la gestión de tareas y paquetes de reconocimiento, además de proveer la capacidad de solicitar dichas tareas a un robot en particular (ver Fig 2.2). Los usuarios de este módulo son dos: el robot, quien solicita y gestiona las tareas de reconocimiento de forma local y remota; y el operador, quien gestiona las tareas y paquetes de reconocimiento en el gestor remoto. Una tarea de reconocimiento es un elemento del sistema de software que ofrece servicios de reconocimiento; y un paquete de reconocimiento, es un paquete de datos que contiene todos los archivos necesarios para ejecutar una tarea de reconocimiento. A continuación se presentan de manera breve la especificación y requerimientos de la estructura general y contenido de los paquetes de reconocimiento.

1. Todos los archivos deben ser compatibles con python 3.
2. La estructura general de un paquete de reconocimiento debe ser la siguiente:

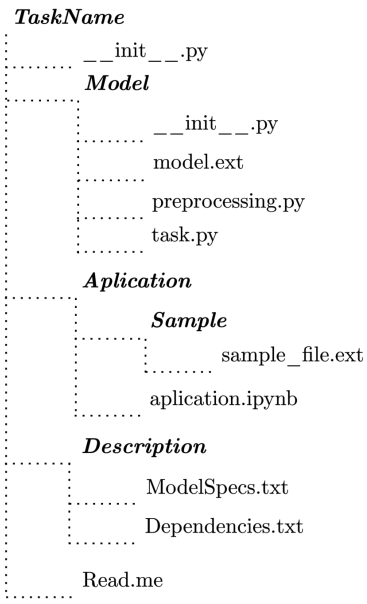


Figura 2.3: *Árbol de directorios de un paquete de reconocimiento*

Los paquetes de reconocimiento deben seguir la estructura que se observa en la Fig. 2.3 para poder ser usados en el módulo reconfigurable de reconocimiento. A continuación se describen todos los elementos que conforman un paquete de reconocimiento:

TaskName: es el directorio principal, éste tiene como nombre el identificador de la tarea de reconocimiento. Está conformado por 3 directorios: Model, Aplication y Description. Adicionalmente, contiene dos archivos: init.py y Read.me.

Model: es un directorio en el cual se almacenan los archivos que contienen el pre-procesamiento de los datos, la versión persistente del modelo o algoritmo entrenado y el archivo de código python correspondiente a la clase task, que permiten integrar los elementos anteriores para realizar las tareas de reconocimiento. Adicionalmente, contiene el archivo: init.py.

Aplication: es un directorio en el cual se genera una pequeña prueba de funcionamiento del algoritmo. Estos archivos se utilizan para que el operador pueda comprobar el funcionamiento del mismo antes de ser

instalado. Para realizar esta prueba, es necesario utilizar un archivo .ipynb, correspondiente a un jupyter notebook; esto con la finalidad de reproducir de manera sencilla las pruebas sobre el algoritmo entrenado. Para realizar estas pruebas es necesario utilizar los archivos contenidos en el directorio Model y además es necesario añadir una o más muestras (directorio Sample), para mostrar detalladamente el pre-procesamiento de la misma en el jupyter notebook.

Description: es un directorio que contiene dos archivos con especificación detallada del algoritmo de reconocimiento y sus dependencias. El archivo ModelSpecs.txt contiene la información de: número de clases, formato de entrada, tipo de salida, nombres de salidas, hiperparámetros, tasa de reconocimiento, entre otros, por otro lado, el archivo Dependencies.txt, contiene las dependencias del paquete de reconocimiento.

- **Módulo afectivo**: este módulo tiene el objetivo de gestionar los estados afectivos del robot en términos de emociones y sentimientos. La entrada de este módulo proviene del módulo cognitivo, el cual le envía la intención de comunicación que ha identificado para que sean actualizados los estados afectivos del robot en función de esa intención, de la persona y de las normas sociales de interacción. Luego, los estados afectivos son enviados a los módulos conductual y cognitivo.

- **Módulo conductual**: el objetivo de este módulo es distribuir en distintos canales o modos, las respuestas y los estados afectivos que serán expresados por el robot. Está constituido por tres componentes: verbal, paraverbal y no verbal. Cada componente se encarga de generar las órdenes respectivas. El componente verbal relacionado con las palabras; el componente paraverbal asociado al volumen de la voz, tono, timbre, etc.; y el componente no verbal relacionado con las expresiones faciales, movimientos, posturas, etc.

2.2 Aprendizaje profundo

El aprendizaje profundo, también conocido en inglés como "deep learning", es una forma de aprendizaje de máquina que permite a las computadoras aprender de la experiencia y comprender el mundo en términos de una jerarquía de conceptos [73], descubriendo una estructura intrincada en grandes conjuntos de datos mediante el uso de un algoritmo de retropropagación para indicar cómo una máquina debe cambiar sus parámetros internos que se utilizan para calcular la representación en cada capa a partir de la representación en la capa anterior [74]. En los últimos años, esta rama de la inteligencia artificial ha logrado excelentes resultados en distintos dominios. Especialmente en el campo de la robótica social y la interacción humano-robot ha sido utilizado en variedad de tareas como seguimiento de la mirada humana [75], análisis de sentimientos a través de la voz [76], reconocimiento de expresiones faciales [77], reconocimiento de palabras claves a través de la voz [50], generación de movimientos para expresar estados del robot [78], entre otras.

2.2.1 Redes neuronales convolucionales (CNN)

En los últimos años, las redes neuronales convolucionales [79] o CNN (por sus siglas en inglés), han demostrado resultados sorprendentes en campos relacionados con el reconocimiento de patrones en imágenes, que van desde la clasificación de imágenes [80] al reconocimiento del habla [81]. Hoy en día, el uso exitoso de las redes neuronales convolucionales en una amplia gama de aplicaciones es una de las razones clave de la creciente popularidad del aprendizaje profundo. Las CNN son un tipo especializado de red neuronal para procesar datos que tiene una topología conocida en forma de cuadrícula [73]. Los ejemplos incluyen datos de series de tiempo, que se pueden considerar como una cuadrícula unidimensional que toma muestras a intervalos de tiempo regulares, y datos de imágenes, que se pueden considerar como una cuadrícula de píxeles de 2 dimensiones. El aspecto más importante de las CNN radica en sus 3 propiedades claves [82]:

1. *Conectividad local*: las imágenes se componen principalmente de elementos (por ejemplo, objetos, animales, texturas, piezas de objetos, etc.). Por lo tanto, las neuronas de una red no necesitan conectarse a todas las unidades de la entrada para encontrar patrones interesantes. En cambio, las neuronas de una red neuronal convolucional solo están conectadas a un pequeño número de unidades en una región espacialmente localizada de la entrada. Esto permite que las neuronas se centren en características locales en lugar de características globales, reduciendo así el número de parámetros drásticamente..

2. *Invarianza espacial*: las imágenes pueden tener elementos en diferentes posiciones sin alterar su contenido semántico, es decir, la red necesita producir valores de salida similares a partir de patrones de entrada similares, independientemente de su ubicación. Las redes neuronales convolucionales implementan esta propiedad al compartir parámetros entre diferentes neuronas.

3. *Características jerárquicas*: los patrones en las imágenes generalmente se pueden descomponer en una jerarquía de características, con características de bajo nivel (por ejemplo, orejas, ojos, nariz) que se pueden agrupar para crear características de alto nivel (por ejemplo, rostros, personas). Al usar múltiples capas, las redes neuronales convolucionales pueden extraer y aprender automáticamente esta jerarquía de características para el reconocimiento de patrones.

Cuando se combinan todas estas propiedades, la red neuronal resultante es altamente eficiente, permitiendo tanto a los investigadores como a los desarrolladores construir modelos más grandes para resolver tareas de reconocimiento con imágenes complejas.

Las redes neuronales convolucionales (ver Fig. 2.4) constan de 3 capas principales: capa convolucional, capa de activación y capa de agrupación. La combinación de esas tres capas es el componente básico de una red neuronal convolucional. Posteriormente les sigue un conjunto de capas completamente conectadas para un razonamiento de alto nivel.

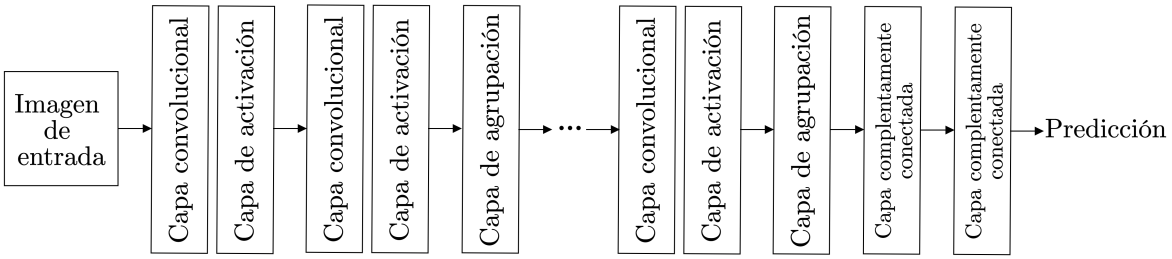


Figura 2.4: *Ejemplo de estructura de una red neuronal convolucional [82]*

2.2.1.1 Capa convolucional

La capa convolucional contiene un conjunto de filtros y su función es realizar una operación de convolución entre estos filtros y la entrada de la capa para crear mapas de características. Un filtro es una cuadrícula de números discretos y, por lo general, tiene forma cuadrada. Sus parámetros, es decir, los números en la cuadrícula, almacenan principalmente un patrón. Este patrón es lo que el filtro detectará en la entrada de la capa (como un detector de características). La operación de convolución es el paso clave que permite que las redes neuronales convolucionales sean invariantes en el espacio. La operación de convolución colocará el filtro sobre la sección superior izquierda de la imagen. Realizará un producto por elementos entre los parámetros del filtro y la cuadrícula correspondiente en la entrada, seguido de la suma del resultado para obtener un valor único (ver Fig. 2.5).

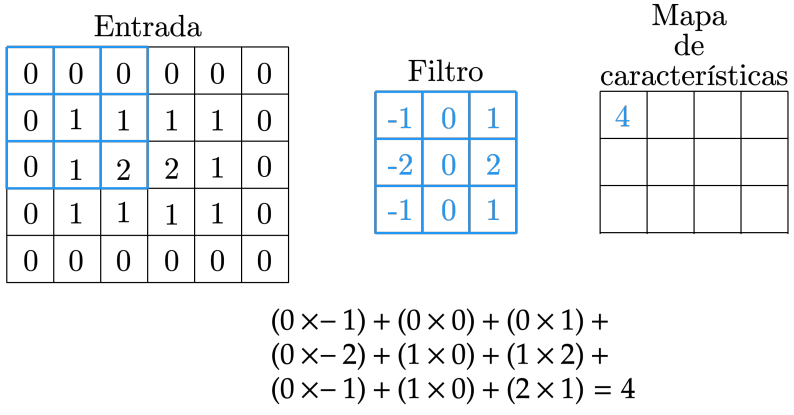


Figura 2.5: *Operación convolución aplicada sobre una entrada*

El valor único resultante indica la presencia o ausencia de la plantilla/patrón único del filtro en esta sección específica de la imagen. A continuación, la operación de convolución deslizará el filtro hacia la derecha y calculará el producto escalar en esta nueva posición. Este deslizamiento de filtro se implementa de izquierda a derecha y de arriba a abajo a través de la entrada y permite la aplicación del filtro en cada posición de la imagen.

Finalmente, el mapa de características almacena el resultado de la operación de convolución. Este almacenamiento se realiza en una estructura de cuadrícula espacial, que mantiene las relaciones espaciales entre la cuadrícula ingresada. Esta propiedad del mapa de características es esencial porque las operaciones de convolución de las siguientes capas dependen fundamentalmente de estas relaciones espaciales. Dado que cada filtro almacena sólo un patrón, es probable que escanear la entrada en busca de un solo patrón resulte en una red muy limitada. Para abordar esta limitación, una capa convolucional debe tener varios filtros, cada uno de los cuales produce un único mapa de características de 2 dimensiones. Después de obtener los diferentes mapas de características, se apilan todos juntos y eso se convierte en el resultado final de la capa convolucional en un volumen de 3 dimensiones con todos los mapas de características. Al igual que en una red neuronal, los parámetros de cada filtro se aprenden durante la fase de entrenamiento. Este procedimiento de aprendizaje implica una inicialización aleatoria de los parámetros del filtro al principio, que luego se sintonizan en muchas iteraciones utilizando un método de descenso de gradiente. De igual manera la capa convolucional tiene diferentes hiperparámetros. Estos son el número de filtros, el tamaño del filtro, tamaño del paso comúnmente llamado “stride” y el relleno también llamado “padding”.

- **Número de filtros**: el número de filtros determina el número de detectores de características. Este hiperparámetro es el más variable entre las capas y, por lo general, se establece en una potencia de 2, entre 32 y 512. El uso de más filtros da como resultado una red neuronal más potente, pero aumenta el riesgo de sobreajuste debido a la mayor cantidad de parámetros a estimar.

- **Tamaño del filtro**: el tamaño (alto y ancho) del filtro define su extensión espacial. Normalmente se utilizan filtros pequeños con rejillas de tamaño 3x3, pero también se usan de 5x5 o 7x7. El uso de filtros pequeños proporciona dos beneficios claves: (1) la cantidad de parámetros que se pueden aprender se reduce significativamente; y (2) asegura que se aprendan patrones distintivos de las regiones locales.

- **Tamaño del paso (stride)**: el tamaño del paso indica el número de píxeles en los que se mueve la ventana del filtro. Su valor suele ser 1, lo que significa que el filtro se desliza píxel a píxel. Sin embargo, se puede aumentar el tamaño del paso si se quiere que el filtro se deslice con un intervalo más grande. Esta alteración hace que el mapa de características resultante sea más pequeño.

- **Relleno (padding)**: a veces es beneficioso rellenar los datos de entrada con píxeles de valor cero alrededor del borde. Esta almohadilla evita que nuestro mapa de características se contraiga durante la operación de convolución porque el píxel central del filtro ahora se puede colocar en el píxel del borde de la imagen de entrada (ver Fig. 2.6). Esto evita un colapso de las dimensiones de las características de salida, lo que permite diseñar redes más profundas.

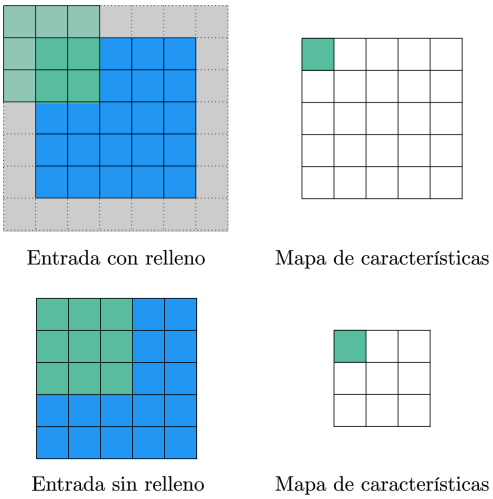


Figura 2.6: *Efecto de aplicar relleno a una entrada*

Desde un punto de vista matemático, la operación de convolución es solo una operación entre dos funciones que se puede expresar de la siguiente manera:

$$y(t) = (x * w)(t) = \int x(\tau)w(t-\tau)d\tau \tag{2.1}$$

Donde $t \in \mathbb{R}$, $\tau \in \mathbb{R}$, $x : \mathbb{R} \to \mathbb{R}$, y $w : \mathbb{R} \to \mathbb{R}$. La función resultante $y : \mathbb{R} \to \mathbb{R}$ después de aplicar el operador de convolución, típicamente denotado con un asterisco $*$, a las funciones x y w es definida como la integral del producto de ambas funciones después de que una se invierte y se desplaza (τ). La primera función x generalmente se denomina entrada, mientras que w es una función de ponderación conocida como kernel. La salida y se denomina mapa de características.

Cuando se implementa la operación de convolución en una computadora, las entradas son discretas y también lo tiene que ser la operación. El índice t solo puede tomar valores enteros. Suponiendo que tanto la entrada como el kernel están definidos solo en t, una convolución discreta se puede definir como:

$$y(t) = (x * w)(t) = \sum_{\tau=-\infty}^{\tau=\infty} x(\tau)w(t-\tau) \tag{2.2}$$

En la práctica, dentro del campo del aprendizaje automático, la entrada y el kernel no son funciones de valor real, sino matrices n-dimensionales de datos con tamaños discretos para cada dimensión. Teniendo todo esto en cuenta, la convolución discreta se puede redefinir como una suma finita sobre matrices n-dimensionales.

$$Y(i,j) \;=\; (X * W)(i,j) = \sum^{m} \sum^{n} X(i+m, j+n)W(m,n) \tag{2.3}$$

2.2.1.2 Capa de activación

Las redes neuronales convolucionales implican el uso de una función de activación no lineal después del cálculo de las operaciones de la capa convolucional. Por lo general, esta función no lineal se define dentro de la capa convolucional. Sin embargo, a veces las transformaciones no lineales se implementan como una capa independiente para permitir una mayor flexibilidad en la arquitectura de la red. Entre las posibles no

linealidades, la función ReLU (ver Fig. 2.7) es la más popular. Matemáticamente la función ReLU se define de la siguiente manera:

$$f(x) = \max(0, x) \tag{2.4}$$

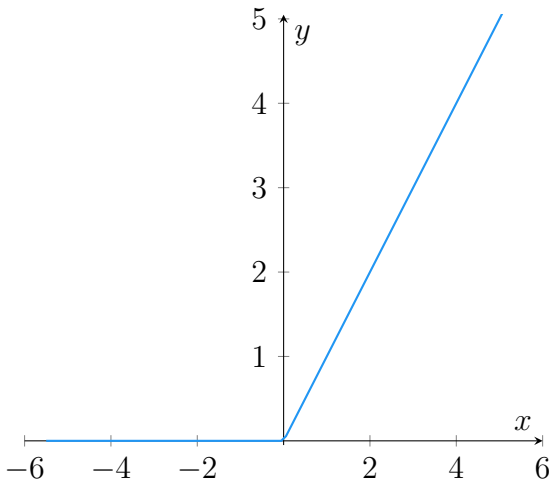


Figura 2.7: *Función ReLU*

El uso de esta función permite entrenar las redes neuronales convolucionales profundas mucho más rápido que el uso de otras funciones de activación, como tangente hiperbólica o sigmoide. La razón es que las funciones tangente hiperbólica y sigmoide se saturan a valores muy altos o muy bajos, haciendo que el gradiente de la función sea muy cercano a cero, lo que ralentiza la optimización del descenso del gradiente. Por otro lado, el gradiente de la función ReLU no está cerca de cero para ningún valor positivo, lo que ayuda a que la optimización converja más rápido.

2.2.1.3 Capa de agrupación

El propósito de la capa de agrupación es reducir el tamaño espacial de la representación capturada por la capa convolucional. Principalmente simplifica la información recopilada y crea una versión condensada de la misma información. La forma más común de agrupación es la agrupación máxima (ver Fig. 2.8). La capa de agrupación máxima desliza una ventana sobre su entrada y toma el valor máximo en la ventana,

descartando todos los demás valores. Similar a una capa convolucional, se deben especificar hiperparametros como el tamaño de la ventana y paso.

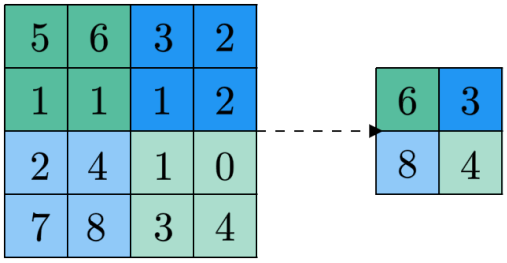


Figura 2.8: *Ejemplo de agrupación máxima*

2.2.1.4 Capa de clasificación

Por lo general, las últimas capas de una red neuronal convolucional son capas completamente conectadas. Su función principal es realizar la clasificación de las características detectadas y extraídas por la serie de capas convolucionales y capas de agrupación. Para ingresar en las capas completamente conectadas, los mapas de características se aplanan en un solo vector unidemensional.

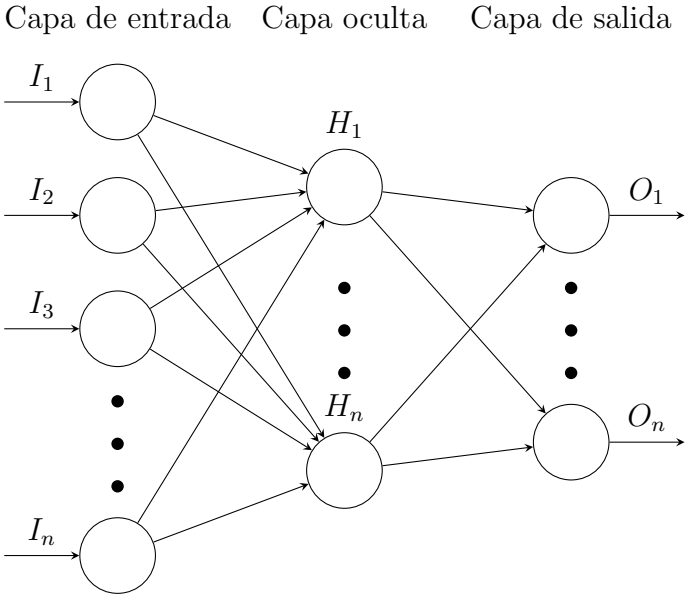


Figura 2.9: *Capas completamente conectadas*

2.2.2 Redes neuronales recurrentes (RNN)

Las redes neuronales recurrentes [83], o RNN (por sus siglas en inglés), son una familia de redes neuronales para procesar datos secuenciales. Así como una red convolucional es una red neuronal especializada para procesar una cuadrícula de valores, como una imagen, una red neuronal recurrente es una red neuronal especializada para procesar una secuencia de valores $x^1, ..., x^\tau$ [73]. Las RNN procesan una secuencia de entrada un elemento a la vez, manteniendo en sus unidades ocultas un vector de estado que contiene implícitamente información sobre la historia de todos los elementos pasados de la secuencia. En los últimos años, las RNN han desempeñado un papel importante en los campos de la visión por computadora [84], el procesamiento del lenguaje natural [85], el reconocimiento del habla [86], entre otros.

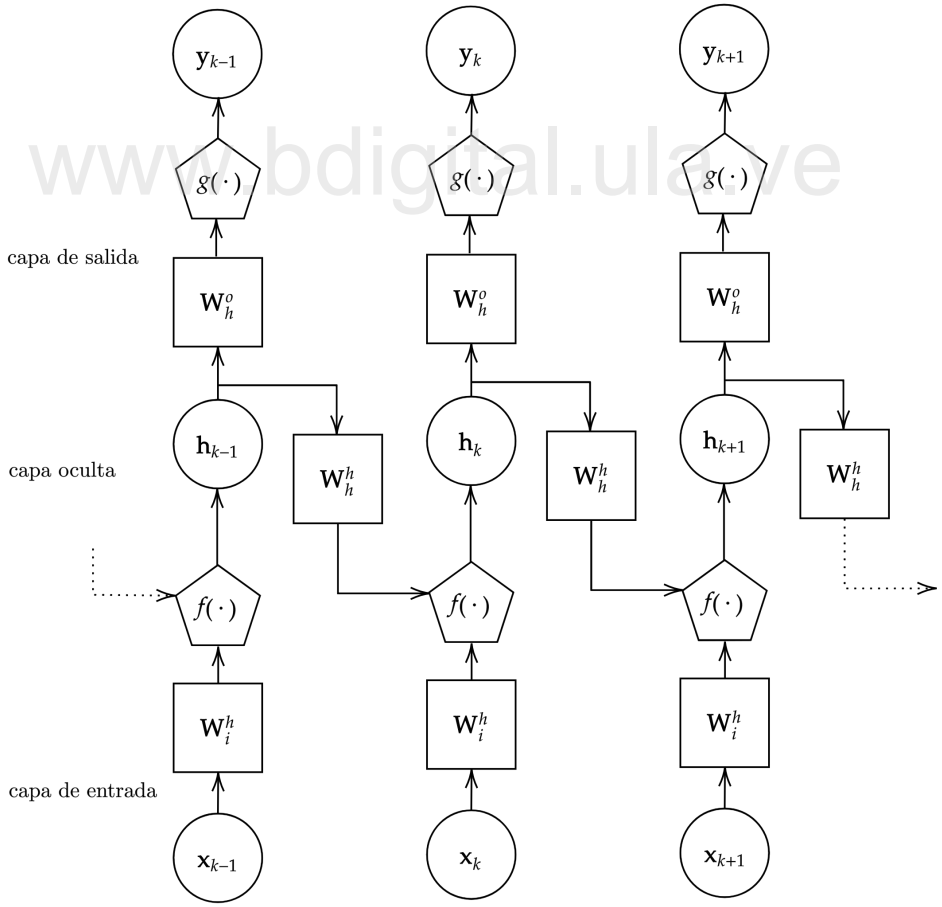


Figura 2.10: *Diagrama de una red neuronal recurrente*

Las capas en una RNN se pueden dividir en capas de entrada, capas ocultas y capas de salida (ver Fig. 2.10). Mientras que las capas de entrada y salida se caracterizan por conexiones de alimentación directa, las capas ocultas contienen conexiones recurrentes. En cada paso de tiempo t, la capa de entrada procesa el componente $\mathbf{x}[t] \in \mathbb{R}^{N_i}$ de una entrada $\mathbf{x}$. La serie de tiempo $\mathbf{x}$ tiene una longitud T y puede contener valores reales, valores discretos, vectores one-hot, etc. En la capa de entrada, cada componente $\mathbf{x}[t]$ se suma con un vector de sesgos $\mathbf{b}_i \in \mathbb{R}^{N_h}$, donde N_h es el número de neuronas en la capa oculta. Luego se multiplica con la matriz de pesos de entrada $\mathbf{W}_i^h \in \mathbb{R}^{N_i \times N_h}$. De manera análoga, el estado interno de la red $\mathbf{h}[t-1] \in \mathbb{R}^{N_h}$ del intervalo de tiempo anterior se suma primero con un vector de sesgo $\mathbf{b}_h \in \mathbb{R}^{N_h}$ y luego se multiplica por la matriz de pesos $\mathbf{W}_h^h \in \mathbb{R}^{N_h \times N_h}$ de las conexiones recurrentes. La entrada actual transformada y el estado pasado de la red son luego combinados y procesados por las neuronas de las capas ocultas, que aplican una transformación no lineal. Las ecuaciones de diferencia para la actualización del estado interno y la salida de la red en un paso de tiempo t son las siguientes:

$$\mathbf{h}[t] = f(\mathbf{W}_i^h(\mathbf{x}[t] + \mathbf{b}_i) + \mathbf{W}_h^h(\mathbf{h}[t-1] + \mathbf{b}_h)) \tag{2.5}$$

$$\mathbf{y}[t] = g(\mathbf{W}_h^o(\mathbf{h}[t] + \mathbf{b}_o)) \tag{2.6}$$

Donde $f(\cdot)$ es la función de activación de las neuronas, generalmente implementada por una función sigmoide o por una tangente hiperbólica. El estado oculto $\mathbf{h}[t]$ transmite el contenido de la memoria de la red en el paso de tiempo t, se inicializa típicamente con un vector de ceros y depende de las entradas pasadas y los estados de la red. La salida $\mathbf{y}[t] \in \mathbb{R}^{N_o}$ se calcula mediante una transformación $g(\cdot)$, generalmente lineal, en la matriz de los pesos de salida $\mathbf{W}_h^o \in \mathbb{R}^{N_r \times N_o}$ aplicada a la suma del estado actual $\mathbf{h}[t]$ y el vector de sesgo $\mathbf{b}_o \in \mathbb{R}^{N_o}$.

2.2.2.1 Celdas de memoria a corto y largo plazo

Uno de los principales inconvenientes de las primeras arquitecturas RNN era su capacidad de memoria limitada, causada por el problema del gradiente que desaparece o explota [87], que se hace evidente cuando la información contenida en entradas pasadas

debe recuperarse después de un intervalo de tiempo prolongado. Las redes neuronales recurrentes de memoria a corto y largo plazo o LSTM (por sus siglas en inglés) [88] son usadas ampliamente hoy en día debido a su rendimiento superior en el modelado preciso de dependencias de datos tanto a corto como a largo plazo. LSTM intenta resolver el problema del gradiente que desaparece o explota al no imponer ningún sesgo hacia las observaciones recientes, pero mantiene el error constante fluyendo hacia atrás en el tiempo. LSTM funciona esencialmente de la misma manera que una red neuronal recurrente común, con la diferencia de que implementa una unidad de procesamiento interno más elaborada llamada *celda* (ver Fig. 2.11). Las diferentes ecuaciones para actualizar el estado de la celda y calcular la salida se enumeran a continuación.

$$\sigma_f[t] = \sigma(\mathbf{W}_f\mathbf{x}[t] + \mathbf{R}_f\mathbf{y}[t-1] + \mathbf{b}_f) \tag{2.7}$$

$$\tilde{\mathbf{h}}[t] = g_1(\mathbf{W}_h\mathbf{x}[t] + \mathbf{R}_h\mathbf{y}[t-1] + \mathbf{b}_h) \tag{2.8}$$

$$\sigma_u[t] = \sigma(\mathbf{W}_u\mathbf{x}[t] + \mathbf{R}_u\mathbf{y}[t-1] + \mathbf{b}_u) \tag{2.9}$$

$$\mathbf{h}[t] = \sigma_u[t] \odot \tilde{\mathbf{h}}[t] + \sigma_f[t] \odot \mathbf{h}[t-1] \tag{2.10}$$

$$\sigma_o[t] = \sigma(\mathbf{W}_o\mathbf{x}[t] + \mathbf{R}_o\mathbf{y}[t-1] + \mathbf{b}_o) \tag{2.11}$$

$$\mathbf{y}[t] = \sigma_o[t] \odot g_2(\mathbf{h}[t]) \tag{2.12}$$

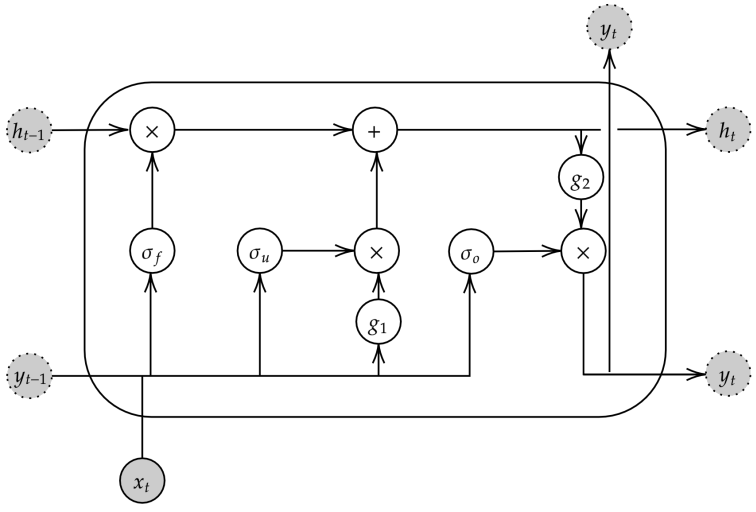


Figura 2.11: *Celda de memoria a corto y largo plazo (LSTM)*

Mientras que una neurona de una RNN implementa una única no linealidad $f(\cdot)$, una celda LSTM está compuesta por 5 componentes no lineales diferentes, que interactúan entre sí de una manera particular. LSTM modifica el estado interno de una celda solo a través de interacciones lineales. Esto permite que la información se propague hacia atrás sin problemas a lo largo del tiempo, con la consiguiente mejora de la capacidad de memoria de la celda. LSTM protege y controla la información en la celda a través de tres puertas, que se implementan mediante una multiplicación sigmoide y puntual. Para controlar el comportamiento de cada puerta, se entrena un conjunto de parámetros con descenso de gradiente, con el fin de resolver una tarea objetivo.

$\mathbf{x}[t]$ es el vector de entrada en el tiempo t. $\mathbf{W}_f$, $\mathbf{W}_h$, $\mathbf{W}_u$ y $\mathbf{W}_o$ son matrices de peso que se aplican a la entrada de la celda LSTM. $\mathbf{R}_f$, $\mathbf{R}_h$, $\mathbf{R}_u$ y $\mathbf{R}_o$ son matrices que definen los pesos de las conexiones recurrentes, mientras que $\mathbf{b}_f$, $\mathbf{b}_h$, $\mathbf{b}_u$ y $\mathbf{b}_o$ son vectores de sesgo. La función $\sigma(\cdot)$ es sigmoide, mientras que $g_1(\cdot)$ y $g_2(\cdot)$ son funciones de activación no lineales puntuales generalmente implementadas como tangentes hiperbólicas. Finalmente, $\odot$ es la multiplicación por entrada entre dos vectores (producto de Hadamard).

Cada puerta de la celda tiene una funcionalidad única y específica. La puerta de olvido σ_f decide qué información debe descartarse del estado de celda anterior $\mathbf{h}[h-1]$. La puerta de entrada σ_u opera en el estado anterior $\mathbf{h}[h-1]$, después de haber sido modificada por la puerta de olvido, y decide cuánto debe actualizarse el nuevo estado $\mathbf{h}[\mathbf{t}]$ con un nuevo candidato $\tilde{\mathbf{h}}[\mathbf{t}]$. Para producir la salida $\mathbf{y}[t]$, primero la celda filtra su estado actual con una no linealidad $g_2(\cdot)$. Luego, la puerta de salida σ_o selecciona la parte del estado que se devolverá como salida. Cada puerta depende de la entrada externa actual $\mathbf{x}[t]$ y la salida de las celdas anteriores $\mathbf{y}[t-1]$. Cuando $\sigma_f = 1$ y $\sigma_u = 0$, el estado actual de una celda se transfiere al siguiente intervalo de tiempo exactamente como está. Con unidades LSTM no ocurre el problema del gradiente que desaparece o explota, debido a la ausencia de funciones de transferencia no lineales aplicadas al estado de la celda. Dado que en este caso la función de transferencia $f(\cdot)$ que se aplica a los estados internos es una función de identidad, la contribución de los estados pasados permanece sin cambios con el tiempo. Sin embargo, en la práctica, las puertas de actualización y olvido nunca están completamente abiertas o cerradas

debido a la forma funcional del sigmoide, que se satura sólo para valores infinitamente grandes. Como resultado, incluso si la memoria a largo plazo en LSTM mejora en gran medida con respecto a las arquitecturas RNN comunes, el contenido de la celda no se puede mantener completamente sin cambios con el tiempo.

2.2.2.2 Unidad recurrente cerrada (GRU)

La Unidad Recurrente Cerrada o GRU (por sus siglas en inglés) [85], es una variación de la celda LSTM que captura de forma adaptativa las dependencias en diferentes escalas de tiempo. En GRU, las puertas de entrada y de olvido se combinan en una única puerta de actualización, que controla de forma adaptativa cuánto puede recordar u olvidar cada unidad oculta. El estado interno en GRU siempre está completamente expuesto en la salida, debido a la falta de un mecanismo de control, como la puerta de salida en LSTM. En una comparación empírica de GRU y LSTM [89], configurados con la misma cantidad de parámetros, se concluyó que, en algunos conjuntos de datos, GRU puede superar a LSTM, tanto en términos de capacidad de generalización como en términos de tiempo necesario para alcanzar la convergencia y actualizar los parámetros.

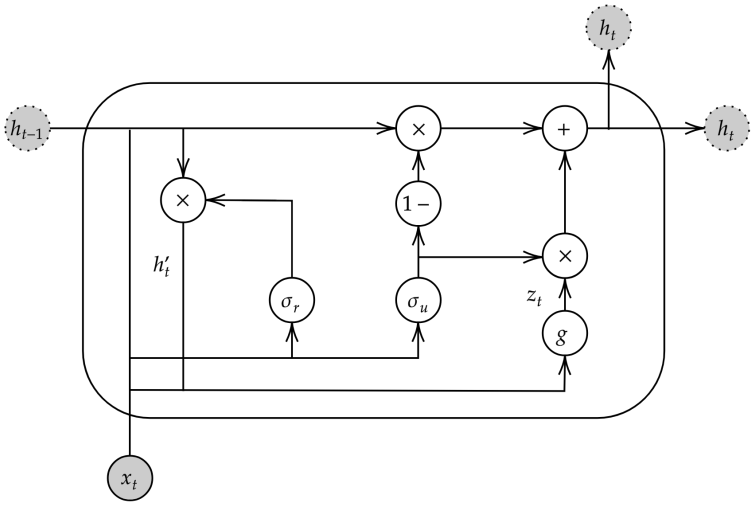


Figura 2.12: *Unidad recucurrente cerrada (GRU)*

En la Fig. 2.12 se muestra una descripción esquemática de la celda GRU. GRU hace uso de dos puertas. La primera es la puerta de actualización, que controla cuánto debe actualizarse el contenido actual de la celda con el nuevo estado candidato. La segunda es la puerta de reinicio que, si está cerrada (valor cercano a 0), puede reiniciar efectivamente la memoria de la celda y hacer que la unidad actúe como si la siguiente entrada procesada fuera la primera en la secuencia. Las ecuaciones de estado del GRU son las siguientes:

$$\mathbf{r}[t] = \sigma(\mathbf{W}_r \mathbf{h}[t-1] + \mathbf{R}_r \mathbf{x}[t] + \mathbf{b}_r) \tag{2.13}$$

$$\mathbf{h}'[t] = \mathbf{h}[t-1] \odot \mathbf{r}[t] \tag{2.14}$$

$$\mathbf{z}[t] = g(\mathbf{W}_z \mathbf{h}'[t-1] + \mathbf{R}_z \mathbf{x}[t] + \mathbf{b}_z) \tag{2.15}$$

$$\mathbf{u}[t] = \sigma(\mathbf{W}_u \mathbf{h}[t-1] + \mathbf{R}_u \mathbf{x}[t] + \mathbf{b}_u) \tag{2.16}$$

$$\mathbf{h}[t] = (1 - \mathbf{u}[t]) \odot \mathbf{h}[t-1] + \mathbf{u}[t] \odot \mathbf{z}[t] \tag{2.17}$$

Aquí, $g(\cdot)$ es una función no lineal generalmente implementada por una tangente hiperbólica. En una celda GRU, el número de parámetros es mayor que en la unidad de una RNN común, pero menor que en una celda LSTM. Los parámetros a aprender son las matrices $\mathbf{W}_r$, $\mathbf{W}_z$, $\mathbf{W}_u$, $\mathbf{R}_r$, $\mathbf{R}_z$, $\mathbf{R}_u$ y los vectores de sesgo $\mathbf{b}_r$, $\mathbf{b}_z$, $\mathbf{b}_u$.

2.2.3 Redes neuronales convolucionales recurrentes (CRNN)

Las redes neuronales convolucionales recurrentes o CRNN (por sus siglas en inglés), son un híbrido de redes neuronales convolucionales (CNN) y redes neuronales recurrentes (RNN). Es tipo de red neuronal está compuesta por varias capas convolucionales seguidas de algunas capas recurrentes. Las CRNN tienen las ventajas de las redes convolucionales y recurrentes. Las capas convolucionales son capaces de extraer de manera eficiente características de nivel medio, abstractas y localmente invariantes de la secuencia de entrada desempeñando así el papel de extractor de características. Las capas recurrentes extraen información contextual de la secuencia de características generada por las capas convolucionales anteriores, lo que permite a este tipo de red neuronal tener en cuenta la estructura global de los datos. Este tipo de red neuronal se

propuso por primera vez en [90] para la clasificación de documentos, posteriormente fue aplicada en otros dominios como la clasificación de imágenes [91] y el reconocimiento del habla [92, 93, 94].

2.3 Reconocimiento de palabras clave

En los últimos años, ha crecido un gran interés en desarrollar aplicaciones orientadas a niños que hacen uso del reconocimiento del habla, por ejemplo, en juegos [53, 95, 96] y robots sociales [48, 50, 57]. Lo que convierte a este tipo de interacción por medio de una interfaz de voz en una modalidad muy deseada por los usuarios infantiles [97]. Sin embargo, solo un pequeño número de estas aplicaciones requieren que se obtenga una transcripción completa de la señal de voz de entrada.

El reconocimiento de palabras clave es una tarea de detección que consiste en descubrir la presencia de palabras habladas específicas en señales de voz [98]. Las aplicaciones de esta tecnología se encuentran generalmente en el contexto de los agentes inteligentes, teléfonos móviles o dispositivos de hogar inteligente [99].

Actualmente, dependiendo de la configuración de la tarea, hay cuatro categorías de enfoques que son la corriente principal para el reconocimiento de palabras clave [52]. El enfoque más básico es simplemente establecer los términos clave en oposición a un modelo de relleno genérico y aplicar una prueba de razón de verosimilitud para identificar las palabras clave. El segundo consiste en realizar el reconocimiento de fonemas (o sílabas u otra unidad de subpalabras) para posteriormente realizar la detección de palabras clave buscando secuencias específicas de fonemas en una grabación y fusionándolas en palabras. La tercera categoría comprende realizar un reconocimiento de vocabulario extenso con un modelo de lenguaje y buscar los términos clave deseados en el sistema de reconocimiento del habla.

En el cuarto conjunto de técnicas, se utilizan ejemplos hablados de las palabras clave para construir detectores de palabras específicos. Nos referimos a estos como métodos basados en ejemplos. Los métodos basados en ejemplos modelan cada palabra clave para ser detectada en su totalidad. Si bien los métodos basados en fonemas son flexibles, los métodos basados en ejemplos son generalmente más precisos o más rápidos. Tales

técnicas basadas en ejemplos también incluyen aquellas basadas en redes neuronales, las cuales en los últimos años, gracias a la creciente popularidad del aprendizaje profundo, se han convertido en el estado del arte de este tipo de sistemas, logrando resultados sorprendentes [92, 99, 100, 101, 102].

Un reconocedor de palabras clave basado en ejemplos consiste en un extractor de características y un clasificador basado en redes neuronales [93] (ver Fig. 2.13). Primero, la señal de voz de entrada de longitud L se enmarca en cuadros superpuestos de longitud l con un tamaño de paso s, dando un total de $T = \frac{L-l}{s} + 1$ fotogramas. De cada trama, se extraen características de voz F, lo que genera un total de características $T \times F$ para toda la señal de voz de entrada de longitud L.

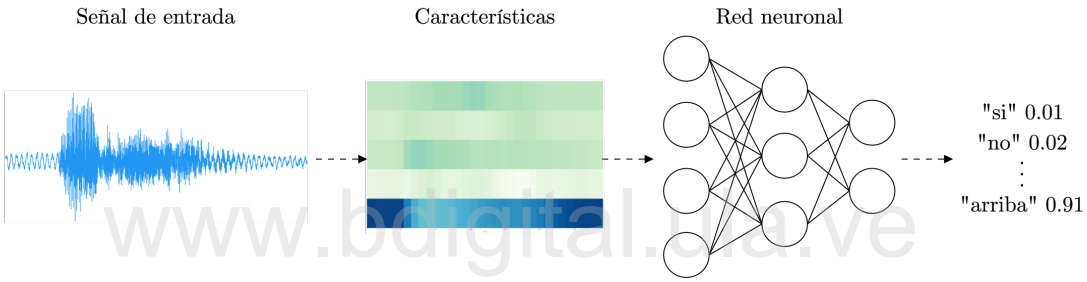


Figura 2.13: *Pipeline de un reconocedor de palabras clave*

Con la matriz de características de voz extraída se alimenta a un módulo clasificador, que genera las probabilidades para las clases de salida. En un escenario del mundo real donde las palabras clave deben identificarse a partir de un flujo de audio continuo, un módulo de manejo posterior promedia las probabilidades de salida de cada clase de salida durante un período de tiempo, mejorando la confianza general de la predicción.

Capítulo 3

Diseño e implementación

3.1 Diseño de la interacción humano-robot para el aprendizaje de las tablas de multiplicación

Como se muestra en la sección de antecedentes, muchas investigaciones han hecho el esfuerzo de eliminar de la mente de los niños que el aprendizaje de las tablas de multiplicación es algo difícil y aburrido, encontrando nuevas estrategias que permiten a los niños aprender las tablas de multiplicación de una manera mucho más fácil y divertida.

Aunque se han obtenido resultados favorables con estas estrategias, se logran destacar principalmente tres desventajas. La primera es que la mayoría de las estrategias solo permiten practicar un conjunto pequeño de las tablas de multiplicación. La segunda consiste en el hecho de que no se proporciona retroalimentación con el resultado correcto cuando los niños se equivocan, evitando así que éstos puedan corregir los errores que cometen. Por último, otra desventaja presente en todas las estrategias consultadas radica en el hecho de que no se adaptan a las necesidades individuales de los niños. Esto representa un problema, puesto que las dificultades matemáticas de los estudiantes varían entre los estudiantes dentro de una clase [103], por lo tanto, cada niño puede presentar dificultades con las tablas de multiplicación distintas a las de sus compañeros, las cuales deben abordarse de manera individual. En resumen, las ventajas y desventajas presentes en estas estrategias se pueden observar en la Tabla 3.1.

Estrategias	Ventajas			Desventajas		
	v1	**v2**	**v3**	**d1**	**d2**	**d3**
Rompecabezas multiplicativo [10]	✓		✓	✓		✓
Cápsulas multiplicativas [10]	✓		✓	✓		✓
Dominó multiplicativo [10]	✓		✓	✓		✓
Juego de la OCA [11]	✓		✓			✓
Bingo de las tablas [12]	✓		✓			✓
Capitán multipli [13]	✓		✓	✓		✓
Sigan la pista [13]	✓		✓	✓		✓
Don Pepe el pescador [13]	✓		✓	✓		✓
Llena la cesta [14]	✓		✓	✓		✓
Tablas de multiplicar [15]	✓	✓			✓	✓
Jugando y cantando voy multiplicando [16]	✓	✓			✓	✓
Multiplication Mat [17]	✓	✓			✓	✓

Tabla 3.1: *Ventajas y desventajas de las estrategias actuales.* ***v1:*** *Basada en juegos,* ***v2:*** *Basada en tecnología,* ***v3:*** *Permite desarrollar habilidades sociales,* ***d1:*** *Sólo permite practicar un conjunto pequeño de las tablas,* ***d2:*** *No hay retroalimentación con el resultado correcto,* ***d3:*** *No se adapta a las necesidades individuales del niño*

En la interacción diseñada, se especificó cómo un robot debe llevar a cabo dos juegos de preguntas y respuestas con un niño. En el primer juego, se le preguntan al niño las operaciones de las tablas de multiplicar del 2 al 9 y éste debe dar la respuesta correcta a cada operación; aquí el robot asume el rol de tutor. En el segundo juego, el robot compite con el niño haciendo preguntas de verdadero y falso con las operaciones de las 3 tablas en las cuales el niño presenta dificultades, esta vez con el robot asumiendo el rol de compañero de juego.

Durante la interacción, cuando el robot pide a un niño que responda una pregunta, éste puede pensar en voz alta mientras intenta responder. El habla de los niños mientras piensan en voz alta es más difícil de reconocer, ya que el volumen del habla es más variado, por otro lado, el niño puede mencionar varias respuestas mientras piensa, lo que dificulta aún más reconocer exactamente cuál de las respuestas que menciona el niño se supone que es la correcta. Para que no se restrinja a los niños la forma en que

dan su respuesta y permitirles pensar en voz alta, y además proporcionar otro tipo de interacciones como por ejemplo solicitar ayuda a un compañero de clases, abordamos este problema siguiendo el patrón de diseño de interacción propuesto en [22] llamado "activación del habla basado en el tacto", para que así los niños puedan proporcionar las respuestas al robot. Es decir, el niño debe presionar un dispositivo táctil cuando sienta que está listo para dar una respuesta, una vez presionado el dispositivo táctil se activará el reconocimiento del habla y el robot escuchará una respuesta durante un período de 4 segundos. A continuación, se explican cada una de las 3 etapas que conforman la interacción diseñada:

Identificación: en esta primera etapa, el robot debe identificar al niño con el que va desarrollar la interacción para que así pueda adaptarse a las necesidades específicas de éste. Una vez identificado, el robot da la bienvenida al niño y lo saluda por su nombre. Por último, dependiendo de si el niño ha interactuado con anterioridad o no, el robot pregunta por la siguiente etapa con la que desea continuar el niño.

Si el niño ya ha interactuado con el robot y ha ejecutado con éxito la etapa de *exploración*, el robot pregunta al niño si desea seguir con la etapa de *exploración* o con la etapa de *aprendizaje*. Si por el contrario, es la primera vez que el niño interactúa con el robot, el robot se presenta a sí mismo y pasa directamente a la etapa de *exploración*. En la Fig. 3.1 se muestra el diagrama de interacción de la etapa de *identificación*.

Exploración: en esta etapa el robot se encarga de aprender las tablas que al niño se le dificultan para que posteriormente puedan ser practicadas en la fase de *aprendizaje*. Una vez que el niño acepta pasar a la etapa de *exploración*, el robot da una introducción al juego de preguntas y respuestas, y explica con detalle las instrucciones que el niño debe seguir. El robot realiza de manera aleatoria 80 preguntas de las tablas de multiplicación del 2, 3, 4, 5, 6, 7, 8 y 9, y por cada respuesta correcta el niño ganará un punto en el juego.

Existen 3 escenarios que deben tomarse en consideración durante el desarrollo del juego. El primero es cuando el niño elige terminar la interacción antes de lo esperado. En nuestra interacción, el niño tiene la opción de detener la interacción mencionando la palabra de activación "*Pepe*" (nombre del robot) seguido del comando "*Detente*", durante el tiempo en el que el robot se encuentra en espera antes de que el niño

presione el dispositivo táctil para dar una respuesta. Cuando el niño decide detener la interacción, el robot responde adecuadamente a la situación a través de una estrategia de "parada suave" como se recomienda en [104], agradeciendo al niño por participar y asegurándose de que el niño no sienta que ha hecho algo malo al elegir retirarse de la interacción. El segundo escenario es cuando el niño necesita que el robot repita nuevamente la pregunta, para este escenario, el niño puede mencionar la palabra de activación "*Pepe*", y luego el comando "*Repite*". El tercer escenario es aquel donde el niño no sabe la respuesta a la operación de la tabla de multiplicación, para este escenario, el niño tiene 2 opciones. La primera opción es mencionar la palabra de activación "*Pepe*", de igual forma que los escenarios anteriores, y luego el comando "*Siguiente*" para pasar a la siguiente pregunta. En la Fig. 3.2 se muestra el diagrama de interacción para este tipo de escenarios. La segunda opción consiste en presionar 2 veces el dispositivo táctil; en el caso de pasar a la siguiente pregunta, el robot marcará la respuesta como incorrecta.

Para motivar a los niños, el robot otorgará recompensas verbales durante el desarrollo de la interacción. Se ha demostrado que factores externos como ofrecer opciones, reconocer los sentimientos de las personas y proporcionar comentarios positivos sobre el desempeño en una tarea mejoran la motivación intrínseca de los estudiantes [105], lo que resulta en un mejor desempeño al momento de realizar una tarea. Escogimos dar una recompensa verbal cada 3 preguntas debido a que aunque los niños con rendimiento alto prefieren un robot con un comportamiento más social, los niños con menor rendimiento pueden distraerse con estas recompensas y desconcentrarse por lo que prefieren un robot con un comportamiento más neutral y menos social [23]; de esta forma pensamos que podemos mantener un equilibrio entre un robot un poco más neutral y un robot social. Las recompensas verbales se han diseñado para que tengan una valencia positiva, indiferentemente de si el niño acierta (ver Tabla 3.2) o se equivoca (ver Tabla 3.3). Una vez completada la etapa, el robot notifica el puntaje que el niño obtuvo en el juego y proporciona una recompensa verbal dependiendo de su rendimiento en el juego: alto (ver Tabla 3.4), medio (ver Tabla 3.5) o bajo (ver Tabla 3.6). Posteriormente, informa las 3 tablas de multiplicación (de menor a mayor) en las que el niño obtuvo el menor rendimiento y se despide del niño.

En la Fig. 3.3 se muestra el diagrama de interacción de la etapa de *exploración*.

Aprendizaje: una vez que el niño haya completado la etapa de *exploración* al menos una vez, puede ejecutar la etapa de *aprendizaje*. En esta etapa el robot desarrolla junto al niño una interacción basada en un juego de preguntas de verdadero o falso, donde el robot pasa de tener el rol de tutor a tener el rol de compañero de juegos. En primer lugar, el robot explica detalladamente las instrucciones que el niño debe seguir. Una vez que inicia el juego, el robot comienza a realizar preguntas del tipo "2 por 2 es igual a 4 ¿verdadero o falso?", a las cuales el niño debe responder siguiendo el mismo mecanismo de "activación del habla basado en el tacto". Las tablas que el robot utiliza para realizar las preguntas son las 3 tablas de multiplicación donde el niño presentó mayores dificultades durante la etapa de *exploración*. Si el niño responde correctamente, obtiene un punto, si falla el robot gana un punto. También se le otorgarán recompensas verbales al niño independientemente si acierta o no cada 3 preguntas, de la misma forma como sucede en la etapa de *exploración*. De igual manera que en la etapa de *exploración*, el niño tiene la oportunidad de detener la interacción, pedir que se repita la pregunta y pasar a la siguiente pregunta durante el desarrollo del juego.

Una vez que se realizan 30 preguntas de verdadero y falso, el robot notifica el ganador del juego e informa el puntaje obtenido tanto por el niño como por el robot. Seguidamente, proporciona una recompensa verbal dependiendo de si el niño ganó o perdió el juego y se despide del niño. En la Fig. 3.4 se muestra el diagrama de interacción de la etapa de *aprendizaje*.

La interacción humano-robot diseñada cubre las desventajas más importantes de las estrategias utilizadas para el aprendizaje de las tablas de multiplicación de la siguiente manera:

- *Sólo permite practicar un conjunto pequeño de las tablas de multiplicar*: los juegos que se desarrollan junto al robot permiten practicar las tablas de multiplicar del 2 al 9.

- *No hay retroalimentación con el resultado correcto*: cuando el niño da un resultado incorrecto el robot proporciona la respuesta correcta.

- *No se adapta a las necesidades individuales del niño*: el robot se adapta a las debilidades del niño con las tablas de multiplicación para practicar las tablas que más se le dificultan.

De igual manera, se toman en consideración las ventajas que poseen estas estrategias:

- *Basada en juegos*: en nuestra interacción el robot desarrolla junto a un niño 2 juegos de preguntas y respuestas.
- *Basada en tecnología*: al ser una interacción humano-robot, debe hacerse uso de un robot social para desarrollar la interacción.
- *Permite desarrollar habilidades sociales*: debido a que la interacción se lleva a cabo con un robot social, lo niños deben interactuar de manera social con un robot, lo que ayuda al desarrollo de habilidades sociales del niño.

Por otro lado, varios de los elementos presentes en las investigaciones de robots sociales en el área de las matemáticas fueron integrados a la interacción, entre éstos se encuentran las recompensas verbales [18, 19] y la capacidad del robot para adaptarse durante la interacción [20, 21, 22], que en nuestro caso, lo logramos adaptando las preguntas que realiza el robot en base a las tablas con las cuales el niño presenta dificultades durante uno de los juegos.

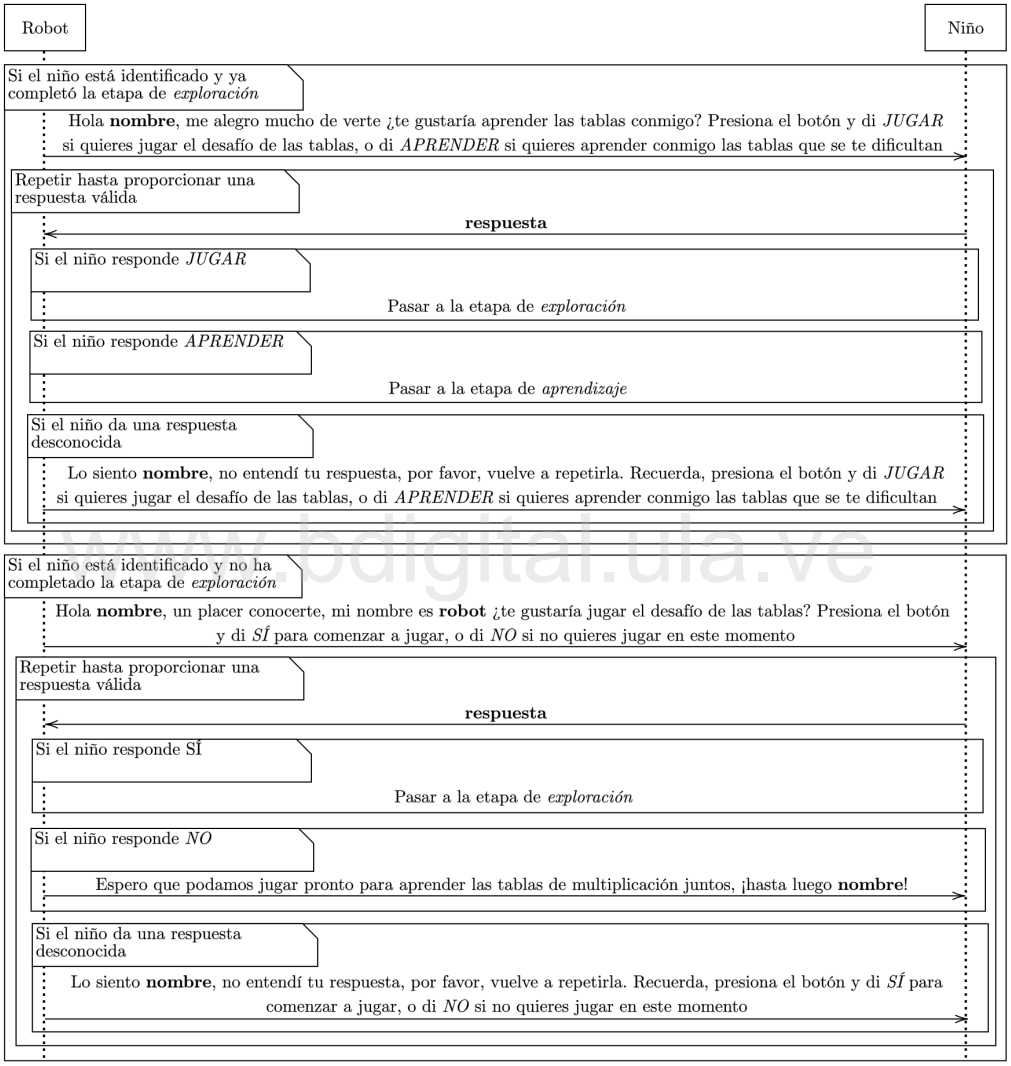


Figura 3.1: *Etapa de identificación.* ***nombre****: nombre del niño,* ***respuesta****: respuestas del niño,* ***robot****: nombre del robot*

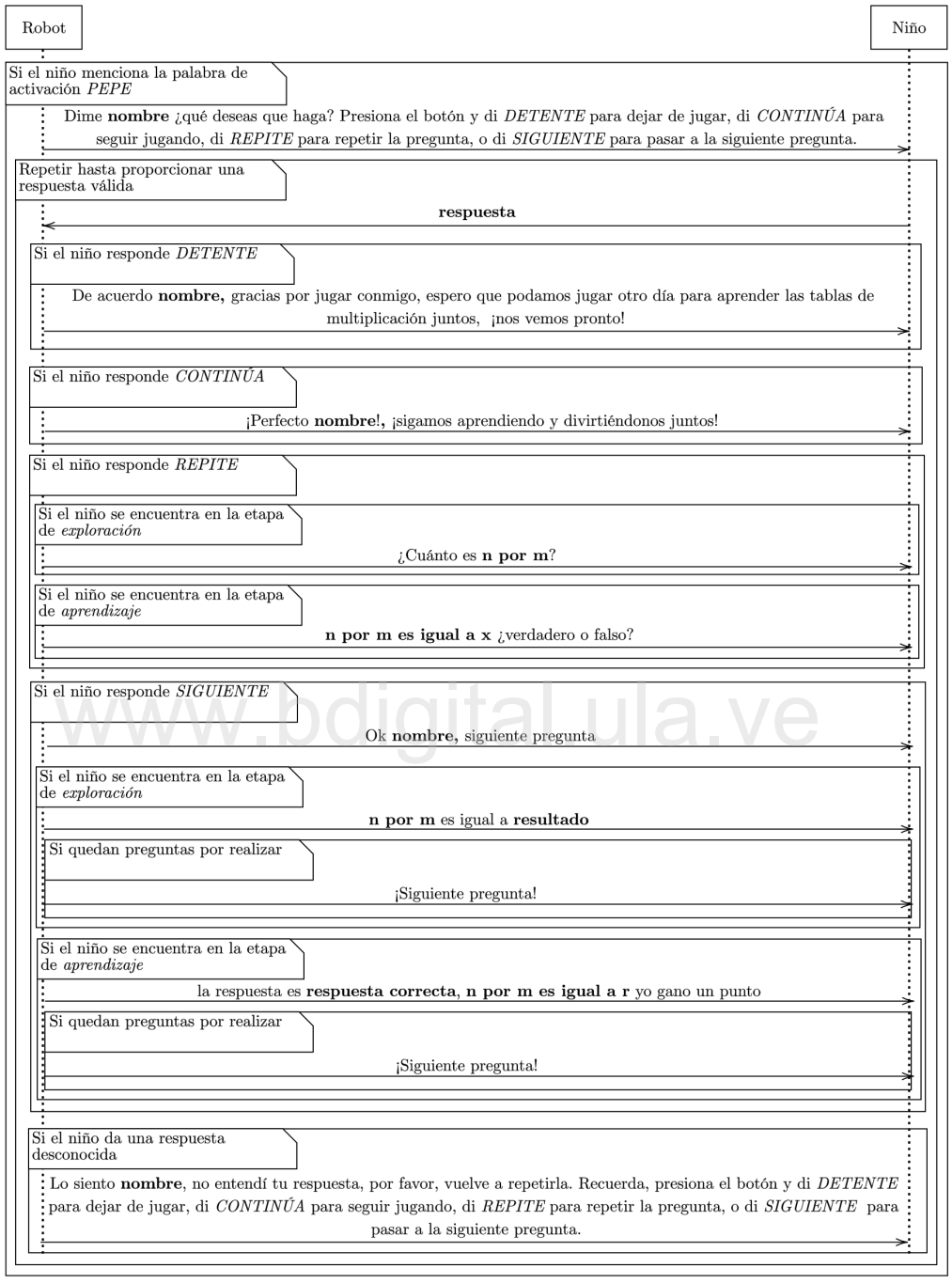


Figura 3.2: *Escenarios para palabra de activación.* ***nombre****: nombre del niño,* ***respuesta****: respuestas del niño, ,* ***n por m****: operación de la tabla de multiplicación etapa de exploración,* ***resultado****: resultado de la operación etapa de exploración,* ***n por m es igual a x****: operación de la tabla de multiplicación propuesta etapa de aprendizaje,* ***respuesta correcta****: respuesta correcta a la pregunta etapa de aprendizaje,* ***n por m es igual a r****: respuesta correcta a la operación propuesta etapa de aprendizaje*

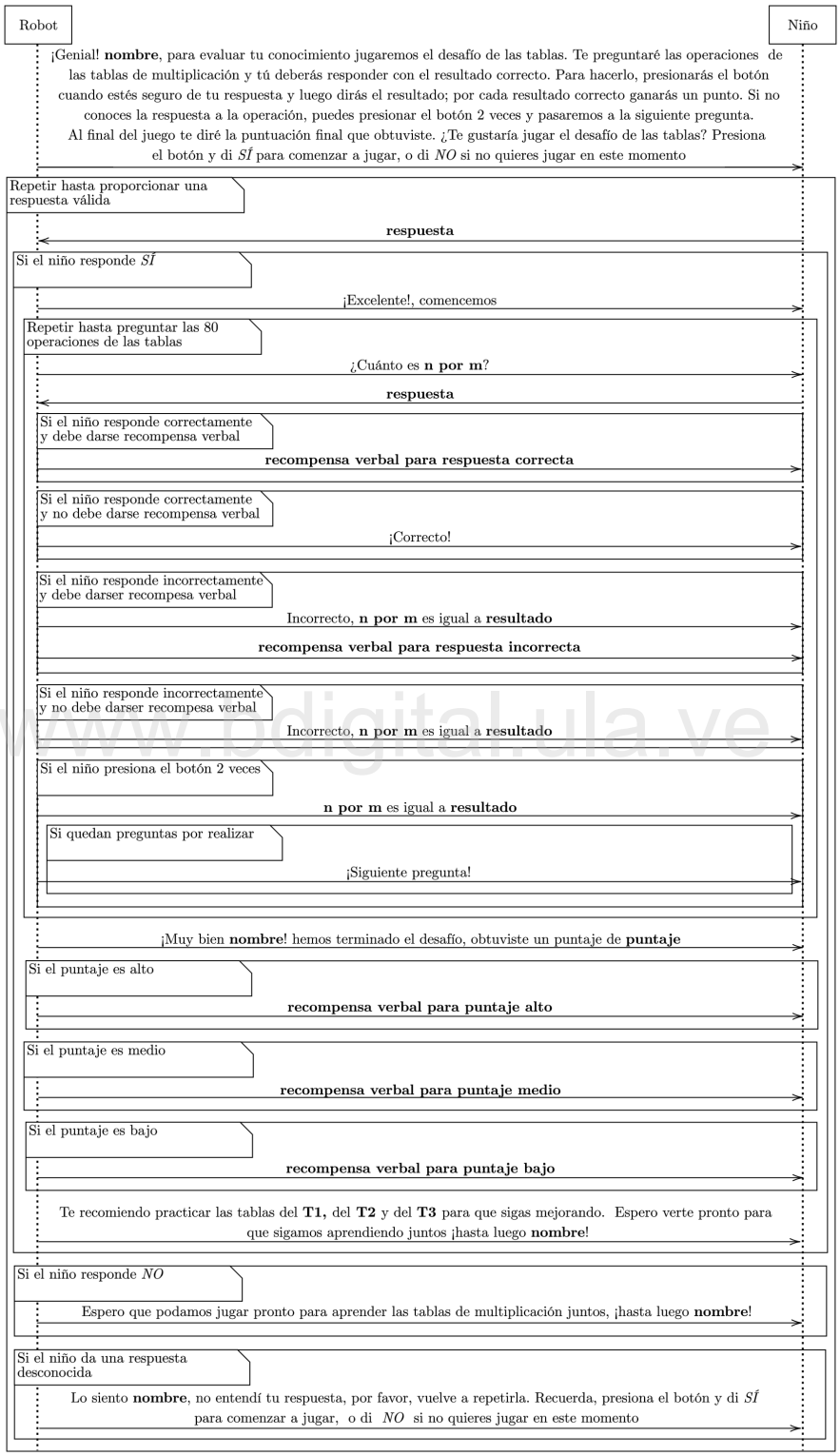


Figura 3.3: *Etapa de exploración.* ***nombre****: nombre del niño,* ***respuesta****: respuestas del niño,* ***n por m****: operación de la tabla de multiplicación,* ***resultado****: resultado de la operación,* ***puntaje****: puntaje obtenido por el niño,* ***Tn****: Tablas de multiplicar que se le dificultan al niño*

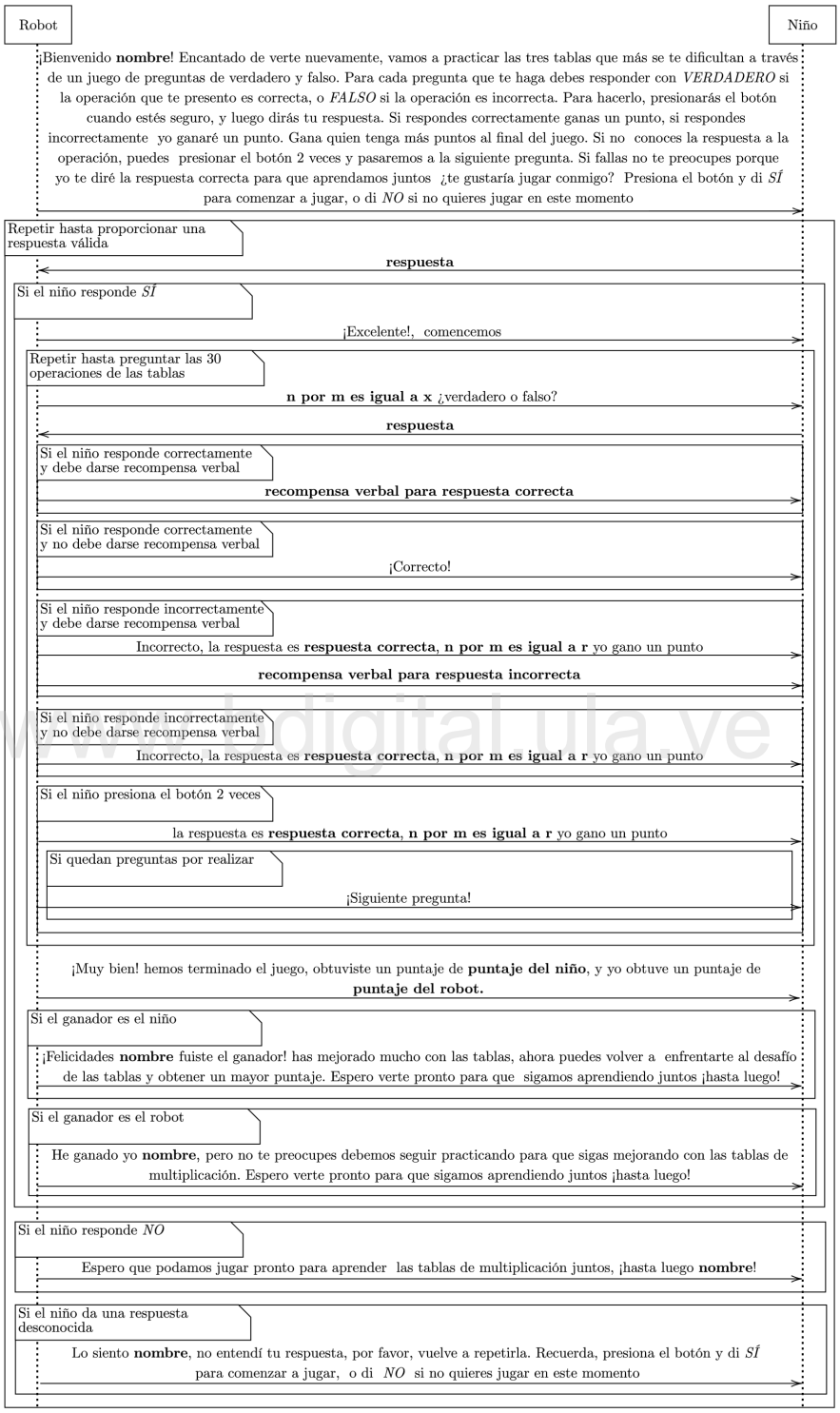


Figura 3.4: *Etapa de aprendizaje.* ***nombre****: nombre del niño,* ***respuesta****: respuestas del niño,* ***n por m es igual a x****: operación de la tabla de multiplicación propuesta,* ***respuesta correcta****: respuesta correcta a la pregunta,* ***n por m es igual a r****: respuesta correcta a la operación propuesta,* ***puntaje del niño****: puntaje obtenido por el niño,* ***puntaje del robot****: puntaje obtenido por el robot.*

Número	Recompensas verbales para respuestas correctas
1	¡Fantástico!, lo estás haciendo muy bien
2	¡Eres un genio **nombre**!, ¡sigue así!
3	¡Asombroso!, esa era la respuesta correcta
4	¡Genial!, vamos muy bien **nombre**
5	¡Increible!, un punto más para tí
6	¡Excelente **nombre**!, ¡sigue así!
7	!Felicidades!, respondiste correctamente!
8	¡Bravo **nombre**!, lo estás haciendo genial
9	¡Enhorabuena!, eso es correcto
10	¡Correcto!, un punto más para tí **nombre**, ¡buen trabajo!
11	¡Buen trabajo! un punto para tí
12	!Fenomenal **nombre**!, un punto más para tí
13	¡Muy bien!, respuesta correcta
14	!Correcto!, sigamos así **nombre**
15	¡Que alegría!, estámos progresando
16	¡Estupendo **nombre**!, ganaste otro punto
17	¡Maravilloso!, la respuesta es correcta
18	¡Esplendido!, lo estás haciendo excelente **nombre**
19	¡Magnífico!, un punto más para tí
20	¡Sigue así **nombre**!, lo estas haciendo bien

Tabla 3.2: *Recompensas verbales para respuestas correctas.* ***nombre****: nombre del niño*

Número	Recompensas verbales para respuestas incorrectas
1	¡No te preocupes **nombre**!, esa era realmente difícil
2	¡No te rindas!, lo vamos a lograr
3	¡Ánimos **nombre**!, ésta era un gran desafío
4	¡Estoy seguro de que la próxima la conseguiremos!
5	¡Tú puedes **nombre**! la próxima vez lo conseguiremos
6	Ésta realmente te hizo pensar, !Sigamos esforzándonos!
7	!Vamos **nombre**!, ¡sí podemos!
8	Ésta es difícil pero lo conseguirás !tú puedes!
9	A la próxima lo conseguirás **nombre**, !Vamos que sí puedes!
10	¡Lo conseguirás para la próxima!, ¡sigue esforzándote!
11	Todo está bien **nombre**, !no te rindas!
12	¡Ánimos!, piensa un poco más las respuestas
13	¡Lo vamos a lograr **nombre**!, pensemos un poco más las respuestas
14	¡Para la próxima sí lo conseguiremos ya verás!
15	!Sigamos trabajando duro **nombre**!
16	!Puedes hacerlo!, para la próxima lo lograremos
17	!La próxima vez lo harás bien **nombre**!, ¡no te preocupes!
18	¡Estoy seguro de que lo vamos a lograr!, piensa un poco más las respuestas
19	¡Yo sé que puedes **nombre**!, ¡no te rindas!
20	Esa era un poco difícil, ¡pero yo sé que tú puedes!

Tabla 3.3: *Recompensas verbales para respuestas incorrectas.* ***nombre****: nombre del niño*

Número	Recompensas verbales para puntaje alto
1	¡Asombroso **nombre**!, ¡eres un maestro de las tablas de multiplicar!
2	!Increíble **nombre**!, ¡eres un genio! dominas muy bien las tablas de multiplicar
3	!Impresionante **nombre**!, lograste un puntaje muy alto en el desafío ¡sigue así!
4	!Fantástico **nombre**!, tienes un muy buen dominio de las tablas de multiplicar
5	!Espléndido **nombre**!, obtuviste un puntaje alto en el desafío, !eres impresionante!

Tabla 3.4: *Recompensas verbales para puntajes alto.* ***nombre****: nombre del niño*

Número	Recompensas verbales para puntaje medio
1	¡Muy bien **nombre**!, Sigue esforzándote, pronto dominarás las tablas de multiplicar
2	¡Lo hiciste bien **nombre**!, debemos seguir practicando para que sigas mejorando
3	¡Nada mal **nombre**!, pronto dominarás por completo las tablas de multiplicar
4	¡Eso estuvo bien **nombre**!, debemos practicar un poco más para que sigas mejorando
5	!Genial **nombre**!, si seguimos practicando lograrás mejorar mucho más

Tabla 3.5: *Recompensas verbales para puntajes medio.* ***nombre****: nombre del niño*

Número	Recompensas verbales para puntaje bajo
1	Sigue practicando **nombre** para que puedas mejorar, ¡tú puedes!
2	Debemos seguir trabajando duro para que puedas mejorar **nombre**, ¡ánimo!
3	No te preocupes **nombre**, para la próxima lo haremos excelente !confío en tí ánimo!
4	Tienes que seguir esforzándote **nombre**, la próxima te irá mejor, !no te rindas!
5	Sigamos practicando juntos y te aseguro que mejorarás, !vamos tú puedes **nombre**!

Tabla 3.6: *Recompensas verbales para puntajes bajo.* ***nombre****: nombre del niño*

3.1.1 Módulos del modelo MIHR considerados

Para organizar y determinar la forma en que se comunican cada uno de los componentes de software que facilitarán el desarrollo de las habilidades sociales del robot al momento de interactuar con el niño durante la estrategia, se toma como base el modelo de interacción humano-robot MIHR. Dentro del modelo MIHR se encuentra el nivel interno del robot (ver Fig. 3.5) encargado de gestionar la dinámica interna de interacción del robot.

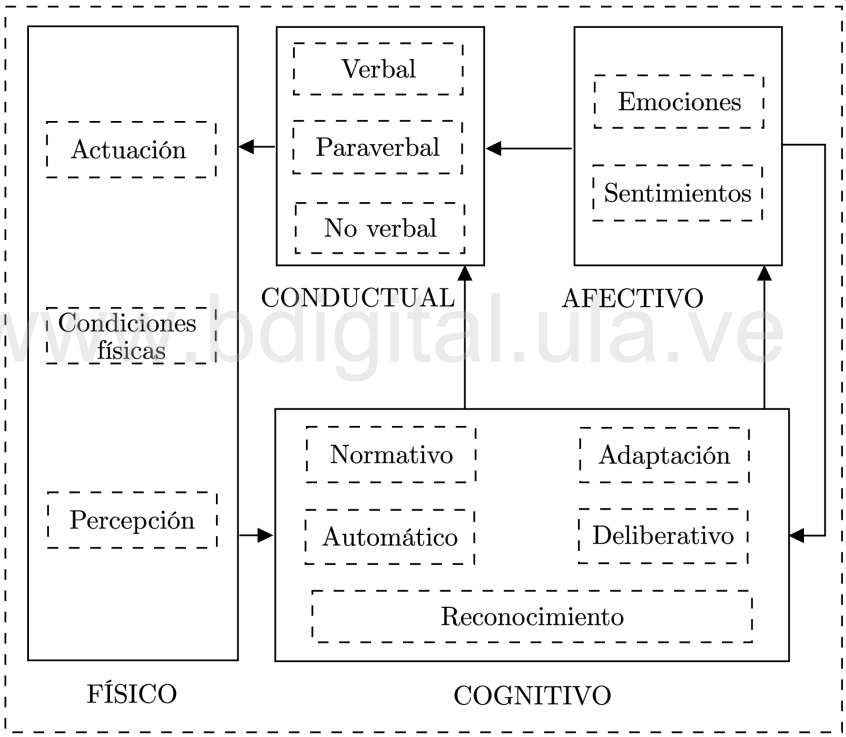


Figura 3.5: *Nivel interno del robot [70]*

No todos los módulos que componen el nivel interno del robot son necesarios para ejecutar la interacción humano-robot diseñada. En la Fig. 3.6, se puede observar cada módulo del cual hacemos uso. A continuación, se describe la forma en que cada módulo deberá ser utilizado.

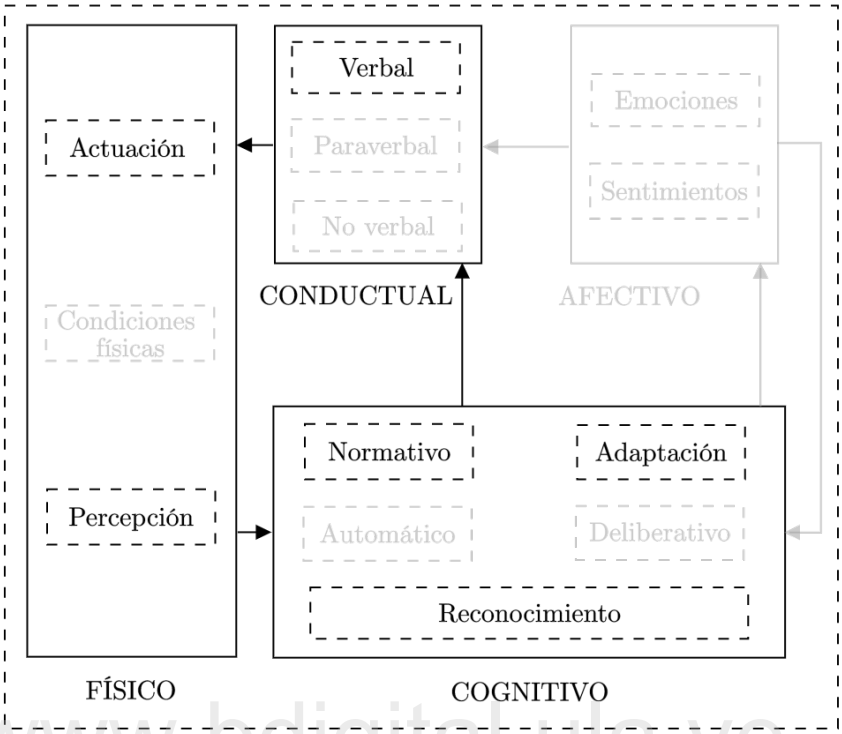


Figura 3.6: *Módulos del nivel interno del robot utilizados*

- **Módulo físico**: del módulo físico, el componente de percepción será el encargado de obtener los datos de entrada que se pasarán al módulo cognitivo. En la interacción humano-robot diseñada, los datos de interés serán las imágenes obtenidas por medio de una cámara para realizar el reconocimiento e identificación de los niños, las señales recibidas a través del dispositivo táctil para proporcionar las respuestas durante la interacción y los audios obtenidos por medio de un micrófono para realizar el reconocimiento del habla. Por otro lado, el componente de actuación se encargará de traducir las órdenes en señales comprendidas por los actuadores, que en este caso serán las bocinas que transmitirán el habla del robot junto con los efectos de sonidos cuando se dicen respuestas correctas o incorrectas.

- **Módulo cognitivo**: de este módulo se usarán los componentes normativo, adaptación y reconocimiento. El componente normativo deberá contener el modelo que describe cada una de las normas que permiten el flujo correcto de la interacción entre el niño y el robot. El componente adaptativo contendrá el modelo que permitirá conocer las 3 tablas que más se le dificultan al niño, para que de esta forma se pueda proporcionar una interacción personalizada. Por último, el componente de reconocimiento se encargará de administrar cada uno de los modelos de reconocimiento necesarios para llevar a cabo la interacción. En nuestro caso, las tareas de reconocimiento son 2: reconocer a los niños a través de imágenes y reconocer números y palabras a través del habla infantil.

- **Módulo conductual**: este módulo será el encargado de manejar la forma en la cual el robot se comunica, en nuestro caso, solo hacemos uso del componente verbal para que el robot se comunique de manera verbal con el niño.

3.2 Diseño y construcción del corpus de audio infantil

En la práctica, la construcción de un corpus de habla infantil representa un reto mayor comparado con los corpus de habla adulta; esto debido a que surgen una serie de desafíos que deben tomarse en consideración. En primer lugar, la capacidad de atención y concentración de los niños depende de su edad [32], lo que conlleva a que los niños puedan distraerse durante sesiones de grabación muy prolongadas dificultando el proceso de grabación. En segundo lugar, los niños pueden presentar dificultades al momento de leer y repetir palabras u oraciones largas o complejas cuando se realizan las grabaciones [8]; esto a causa de que la producción del habla es una actividad motora compleja que los niños todavía están aprendiendo a dominar. Por lo tanto, es necesario utilizar recursos adicionales como diapositivas, animaciones, descansos durante las grabaciones y reproducciones de audio de las sentencias a pronunciar que sirvan de guía a los niños para así contrarrestar estos problemas.

Los corpus de audio infantil pueden ser clasificados en 2 tipos según la forma en la que se solicite a los niños expresar el discurso a grabar durante el proceso de grabación.

- **Discurso espontáneo**: son aquellos donde el discurso expresado por el niño se obtiene por medio de algún tipo de narración o discursos provocados de manera natural.
- **Discurso leído**: son aquellos donde el discurso expresado por el niño se obtiene por medio de la lectura de las declaraciones de interés.

Dentro de los corpus de audio infantil de discurso espontáneo podemos encontrar el corpus de audio "NITK Kids' Speech Corpus" [33]. De discurso leído "TBALL" [32], "CHOREC" [34], "CNG" [8] y "CID children's speech corpus" [35]. También existe el caso donde el corpus de audio contiene grabaciones tanto de discurso espontáneo como de discurso leído, siendo éste el caso del corpus "OGI Kids' Speech" [36].

En este proyecto de grado se construirá un corpus de audio infantil de discurso leído, que llevará por nombre "LaSDAI Comandos de Voz Infantil" (LaSDAICVI). El cual estará destinado al entrenamiento y evaluación de modelos de reconocimiento de palabras clave en español a través del habla infantil, y que hasta el momento de realización de este proyecto de grado, sería el primer corpus de audio infantil en español para este tipo de aplicaciones.

3.2.1 Participantes

El corpus de audio LaSDAICVI fue recolectado de un total de 41 niños inscritos en escuelas primarias, pertenecientes a los grados tercero a sexto con edades comprendidas entre los 8 y 11 años ($\mu = 9.609756098$, $\sigma = 1.069533748$). Para cada niño que formó parte de las grabaciones, los padres dieron su consentimiento a través de la firma de un consentimiento informado para permitirles participar y proporcionaron información relevante como el nombre, género, edad y grado. De igual forma, a todos los niños que estuvieron de acuerdo en participar se les pidió que firmaran un asentimiento informado. En la Tabla. 3.7 se puede observar la cantidad de niños por grado y género que participaron en las grabaciones.

Grado	Femenino	Masculino	Total
3er	5	5	10
4to	6	5	11
5to	7	5	12
6to	4	4	8

Tabla 3.7: *Cantidad de niños por grado y género*

3.2.2 Palabras y números grabados

Las palabras y números grabados consistieron en la serie de números del 0 al 9, junto con los números resultantes en las operaciones de las tablas de multiplicación del 2 al 9, además de 18 palabras necesarias que servirán como comandos de voz para desarrollar la interacción diseñada. Las palabras seleccionados fueron: “Pepe”, “Sí”, “No”, “Detente”, “Continúa”, “Repite”, “Atrás” “Siguiente”, “Apágate”, “Actívate”, “Arriba”, “Abajo”, “Izquierda”, “Derecha”, “Jugar”, “Aprender”, “Verdadero” y “Falso”. Las palabras seleccionadas fueron inspiradas por el corpus de audio adulto Speech Commands [106], el cual se ha convertido en uno de los corpus de audio de habla adulta más usados para el entrenamiento y evaluación de reconocedores de palabras clave.

3.2.3 Equipo de grabación

El habla de los niños fue grabada a una frecuencia de muestreo de 16000 Hz, 32 bits de resolución y utilizando un solo canal a través de un micrófono de condensador. El micrófono de condensador fue conectado a un computador portátil donde se ejecutaba el software de edición y grabación de audio Audacity para realizar las grabaciones. Un segundo computador portátil se conectó a un monitor para presentar las diapositivas que contenían las palabras y números, junto con una reproducción en audio de la misma en sincronía con una animación 2D de un robot (ver Fig. 3.7), las cuales eran controladas por el experimentador encargado de las grabaciones. La reproducción de audio (de las palabras y números) era escuchada por el niño a través de unos audífonos auriculares conectados al computador portátil para evitar que al momento

que ésta se reprodujera interfiriera con la captura de audio. En la Fig. 3.8 se muestra la configuración usada para realizar las grabaciones.

Figura 3.7: *Diapositiva con animación 2D del robot Pepe*

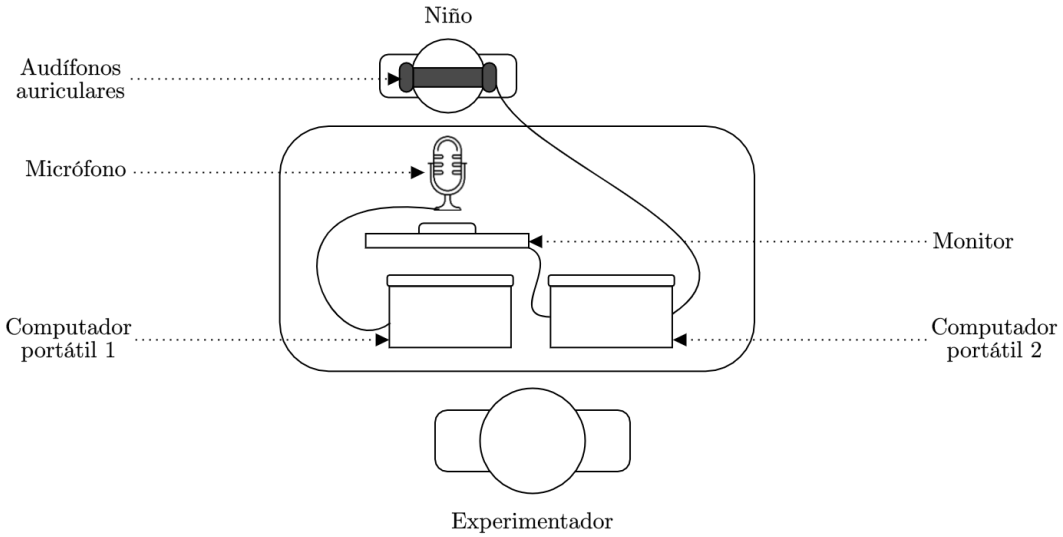


Figura 3.8: *Configuración usada para realizar las grabaciones*

3.2.4 Protocolo de grabación

El proceso de grabación fue realizado en una habitación con poco ruido. Antes de comenzar las grabaciones se le explicaba al niño en qué consistía la sesión de grabación, se le pedía que firmara el asentimiento informado y se le asignaba un identificador único. Luego se realizaba una pequeña sesión de práctica donde se grababan 5 palabras y los números del 0 al 9; esto con la intención de que los niños que estuvieran nerviosos se sintieran más familiarizados con el proceso de grabación. Para obtener la mayor cantidad de muestras por niño, se realizaron dos sesiones de grabación cada una con una duración máxima de 25 minutos. En cada sesión, los niños debían repetir la serie de palabras y números un total de 5 veces, para así obtener un mínimo de 10 muestras por niño para cada palabra y número. El proceso de grabación fue realizado como se explica a continuación:

1. A cada uno de los niños se les pidió sentarse en una silla ubicada a 30 cm del micrófono.

2. A cada niño se le pedía que pronunciara el número o palabra que se mostraba en el monitor luego de la reproducción en audio de la misma. De esta forma evitábamos que los niños cometieran una mayor cantidad de errores durante las grabaciones.

3. Se presentaba la serie de números de forma aleatoria hasta completar las 5 repeticiones para cada número, tomando breves descansos cada 42 números para evitar la fatiga en los niños.

4. Se presentaban las palabras hasta completar las 5 repeticiones para cada una, tomando breves descansos cada 18 palabras presentadas.

5. En caso de pronunciar incorrectamente alguna palabra o número, capturar algún ruido fuerte del exterior o que simplemente el niño o el experimentador no estuviese satisfecho con el resultado, se le pedía nuevamente repetir la palabra o número.

6. Una vez finalizada la primera sesión de grabación se les otorgó a los niños un descanso de 10 minutos donde se les brindaba un refrigerio. Luego de haber terminado el descanso se les pidió repetir el proceso de grabación para obtener las muestras restantes.

3.2.5 Etiquetado de las grabaciones

Cada una de las grabaciones fue etiquetada y recortada para extraer las palabras y números cuidadosamente de manera manual por un experimentador, utilizando el software de edición de audio Audacity, descartando aquellas palabras o números de poca calidad (pronunciaciones ambiguas o presencia de ruidos fuertes). Seguidamente, cada palabra y número fue almacenado en formato WAV 16bits PCM, y se renombraron según la información obtenida de los niños usando la siguiente convención de nombre:

$$\{ID\}\_\{EDAD\}\_\{GÉNERO\}\_\{GRADO\}\_\{\#GRABACIÓN\}.\text{wav}$$

- $\{ID\}$: identificador del niño.
- $\{EDAD\}$: edad del niño.
- $\{GÉNERO\}$: género del niño.
- $\{GRADO\}$: grado que estudia el niño.
- $\{\#GRABACIÓN\}$: número de la muestra grabada.

Posteriormente, se guardaron en una carpeta etiquetada con la palabra o número presente en la grabación. El corpus de audio final consistió en 29061 muestras de audios. En la Tabla. 3.8 se muestra el número total de muestras obtenidas para cada palabra y número.

Tabla 3.8: *Número de muestras por palabra y número en el corpus de audio infantil LaSDAICVI*

Palabra o número	**Número de muestras**
Pepe	467
Sí	469
No	461
Verdadero	486
Falso	478
Detente	472
Continúa	487
Siguiente	462
Atrás	457
Apágate	465
Actívate	486
Arriba	481
Abajo	466
Izquierda	481
Derecha	461
Aprender	501
Jugar	503
Repite	508
0	504
1	482
2	481
3	481
4	479
5	493
6	483
7	472
8	497
9	472
10	444
12	472
14	505
Continua en la próxima página	

Tabla 3.8 – continuación de la página previa

Palabra o número	Número de muestras
15	494
16	487
18	491
20	474
21	469
24	492
25	499
27	501
28	495
30	490
32	506
35	504
36	498
40	475
42	483
45	475
48	463
49	478
50	466
54	494
56	491
60	498
63	486
64	500
70	495
72	537
80	483
81	494
90	487

3.3 Diseño e implementación de los modelos de reconocimiento de habla infantil

Uno de los mayores desafíos técnicos presentes en la interacción humano-robot, y especialmente en la interacción de robots con niños, es la capacidad de percepción del robot [107]. Por lo general, se espera que un robot pueda percibir su entorno de la misma manera que lo hace un humano. Sin embargo, recrear artificialmente ese nivel de percepción es una tarea muy complicada.

Un ejemplo de esto es el reconocimiento del habla, porque, aunque el reconocimiento de habla ha logrado grandes avances en los últimos años, el reconocimiento del habla infantil en escenarios de interacción humano-robot todavía tiene un rendimiento inferior [57]. Al practicar las tablas de multiplicar, no se necesita mucho lenguaje verbal y la tarea es relativamente simple en interacción con un robot social. Esto permite el uso del habla de manera limitada, logrando una interacción mucho más cercana a la parecida con un humano, mientras que la mayoría de las interacciones con robots dependen de las tabletas que las acompañan o de técnicas como la del "Mago de Oz". Para lograr la comunicación verbal entre el niño y el robot escogimos un enfoque de reconocimiento de palabras clave como en [50], para determinar qué palabras o números son pronunciados por un niño durante la interacción.

El principal objetivo del reconocimiento de palabras clave es detectar un conjunto relativamente pequeño de palabras predefinidas, en un flujo de audio dicho por un usuario, siendo éste usado generalmente en el contexto de un agente inteligente, un teléfono móvil o un dispositivo de hogar inteligente. Por lo general, este tipo de tecnología se aplica a dominios en los que un reconocedor del habla completo es difícil de desarrollar e innecesario. En particular, en el campo de la robótica es utilizada para permitir a las personas controlar a los robots a través de comandos de voz para que estos realicen alguna acción en específico [50, 108].

Normalmente, un reconocedor de habla completo se ejecuta en la nube haciendo uso de servidores con grandes capacidades computacionales, esto requiere la transferencia de grabaciones de audio desde el dispositivo del usuario hasta los servidores en la nube, existiendo así importantes implicaciones de privacidad. Un reconocedor de palabras

clave puede ejecutarse directamente desde el dispositivo, lo que permite abordar 3 limitaciones clave: en primer lugar, el reconocimiento de comandos comunes como "Encendido" y "Apagado", así como otras palabras frecuentes como "Sí" y "No", se puede lograr directamente en el dispositivo del usuario evitando así cualquier posible problema de privacidad, lo cual, al trabajar con niños, es un factor muy importante a tomar en consideración. En segundo lugar, al realizar el reconocimiento de palabras clave desde el dispositivo se obtiene una respuesta con una latencia mínima ya que no hay ida y vuelta con un servidor. Por último, al realizar el reconocimiento de palabras clave desde el dispositivo, no se requiere de una conexión a internet.

3.3.1 Lista de palabras clave para cada modelo

El alcance de este proyecto de grado es reconocer a través del habla los números que se encuentran como resultado en las operaciones de las tablas de multiplicación del 2 al 9, así como las palabras requeridas durante la interacción diseñada para el aprendizaje de las tablas de multiplicación (ver sección 3.1).

Para lograr este objetivo y abordar el problema de una forma más eficiente, se optó por dividirlo en problemas de menor complejidad. Por lo tanto decidimos crear un total de 10 modelos de reconocimiento de palabras clave: 1 modelo para cada tabla de multiplicación que reconozca los números presentes en los resultados de la tabla (para un total de 8 modelos); 1 modelo para reconocer los comandos y palabras requeridas durante la interacción; y 1 modelo para reconocer la palabra de activación "Pepe".

De igual manera, cada uno de los modelos debe reconocer cuando hay presencia de ruido/silencio (_silencio_) y cuando se pronuncia una palabra o número desconocido para cada modelo (_desconocido_). Para las palabras o números desconocidos, decidimos agregar como números desconocidos aquellos números que tengan una pronunciación similar a los números que se deben reconocer para cada modelo de tabla de multiplicar. Para los modelos de *interacción* y *activación*, seleccionamos algunas palabras y números del conjunto de datos LaSDAICVI para que sirvieran como palabras o números desconocidos. El uso de palabras o números desconocidos ayudará a reducir la tasa de falsos positivos en los modelos. En la Tabla. 3.9 se observa las palabras clave objetivo a reconocer para cada uno de los modelos.

Modelo	Palabras clave objetivo	Palabras o números desconocidos
Tabla del 2	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, _silencio_, _desconocido_	15, 21, 24, 25, 27, 28, 30, 32, 36, 40, 42, 48, 50, 54, 56, 60, 64, 79, 72, 80, 90
Tabla del 3	3, 6, 9, 12, 15, 18, 21, 24, 27, 30, _silencio_, _desconocido_	1, 2, 4, 7, 8, 10, 14, 16, 20, 25, 28, 32, 35, 36, 40, 48, 49, 50, 54, 56, 60, 63, 64, 70, 80, 90
Tabla del 4	4, 8, 12, 16, 20, 24, 28, 32, 36, 40, _silencio_, _desconocido_	2, 6, 10, 14, 15, 18, 21, 25, 27, 30, 35, 42, 45, 48, 49, 50, 54, 56, 60, 64, 70, 72, 80, 90
Tabla del 5	5, 10, 15, 20, 25, 30, 35, 40, 45, 50, _silencio_, _desconocido_	12, 14, 16, 18, 21, 24, 27, 28, 32, 36, 42, 48, 49, 54, 56, 60, 70, 80, 90
Tabla del 6	6, 12, 18, 24, 30, 36, 42, 48, 54, 60, _silencio_, _desconocido_	2, 4, 8, 10, 14, 15, 16, 20, 21, 25, 27, 28, 32, 35, 40, 45, 49, 50, 56, 63, 64, 70, 80, 90
Tabla del 7	7, 14, 21, 28, 35, 42, 49, 56, 63, 70, _silencio_, _desconocido_	1, 2, 3, 5, 6, 8, 9, 12, 15, 16, 18, 20, 24, 25, 27, 30, 32, 36, 40, 45, 48, 50, 54, 60, 64, 72, 80, 81, 90
Tabla del 8	8, 16, 24, 32, 40, 48, 56, 64, 72, 80, _silencio_, _desconocido_	2, 4, 6, 10, 18, 20, 21, 25, 27, 28, 30, 35, 36, 42, 45, 49, 50, 54, 60, 63, 70, 81, 90
Tabla del 9	9, 18, 27, 36, 45, 54, 63, 72, 81, 90, _silencio_, _desconocido_	1, 2, 3, 4, 5, 6, 7, 8, 10, 16, 20, 21, 24, 25, 28, 30, 32, 35, 40, 42, 48, 49, 50, 56, 60, 64, 70, 80
Interacción	sí, no, verdadero, falso, detente, siguiente, aprender, repite, jugar, continúa, _silencio_, _desconocido_	pepe, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Activación	pepe, _silencio_, _desconocido_	sí, no, verdadero, falso, detente, siguiente, aprender, repite, jugar, continúa, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9

Tabla 3.9: *Palabras clave objetivo y palabras clave desconocidas para cada modelo*

3.3.2 División del corpus de audio LaSDAICVI

Para crear los conjuntos de entrenamiento, validación y prueba, se dividió el corpus de audio LaSDAICVI de la siguiente manera: por cada grado y género, realizamos un muestreo sin reemplazo para seleccionar de manera aleatoria el 50% de los niños para el conjunto de entrenamiento, 25% para el conjunto de validación y 25% para el conjunto de prueba. De esta forma mantenemos los niños por grado y género equilibrados dentro de cada conjunto de datos. Al mismo tiempo, evitamos que un niño aparezca en diferentes conjuntos de datos, previniendo así la fuga de datos.

En total, se obtuvo un conjunto de entrenamiento con 21 niños y 14737 muestras de audio; un conjunto de validación con 10 niños y 7210 de muestras de audio; y un conjunto de prueba con 10 niños y 7113 muestras de audio. A continuación se muestran los identificadores, edad, género y grado de cada niño perteneciente a cada conjunto de datos.

#	**Identificador**	**Género**	**Edad**	**Grado**
1	2d5ywQE8	F	9	3ro
2	etVCpoHz	F	8	3ro
3	68QfMqdS	F	8	3ro
4	JE8LZ4so	M	8	3ro
5	SHjKaKWT	M	8	3ro
6	eJcUaRSW	M	8	3ro
7	MwPvEkhC	F	9	4to
8	KEqDpnec	F	10	4to
9	HGhtQEgj	M	10	4to
10	fzbcST9H	M	10	4to
11	GrLJrJfs	M	9	4to
12	UCseQ6gc	F	10	5to
13	QsoNpuJa	F	11	5to
14	e4PAirZ2	F	10	5to
15	8spkKHdE	M	10	5to
16	WpG4B6T3	M	10	5to
17	BH25qTuQ	M	10	5to
18	7DMGwJNa	F	11	6to
19	S6qS4hD9	F	11	6to
20	ijhfhsjo	M	11	6to
21	5j7Uj7vn	M	11	6to

Tabla 3.10: *Niños pertenecientes al conjunto de entrenamiento*

#	Identificador	Género	Edad	Grado
1	Mvx26dQ7	F	8	3ro
2	fzm5ARkQ	M	8	3ro
3	EWaAuiQt	F	10	4to
4	ZD24aEoE	F	9	4to
5	FzJCbWZY	M	9	4to
6	ctd7zjAm	F	10	5to
7	e5TFkGRp	F	10	5to
8	k38ovYSi	M	10	5to
9	XAZZ8qWp	F	11	6to
10	PNF7NjoD	M	11	6to

Tabla 3.11: *Niños pertenecientes al conjunto de validación*

#	Identificador	Género	Edad	Grado
1	Wf4zy8ui	F	8	3ro
2	mYw3UQoH	M	8	3ro
3	Nj5XnFzn	F	9	4to
4	HcFUSEdx	F	9	4to
5	77fYRs5a	M	10	4to
6	Zx9SYM4g	F	10	5to
7	HuwJU54m	F	10	5to
8	WEddoaXU	M	10	5to
9	CLTgXEu9	F	11	6to
10	QZP6LcEN	M	11	6to

Tabla 3.12: *Niños pertenecientes al conjunto de prueba*

Una vez dividido el corpus de audio, se seleccionaron específicamente las palabras clave a reconocer por cada modelo desde cada conjunto de datos para formar los conjuntos de entrenamiento, validación y prueba individuales para cada modelo. Debido a que la palabra clave objetivo _desconocido_ está conformada por varias palabras o números del corpus de audio, ésta posee un mayor número de muestras en relación con las otras palabras clave objetivo. Para mantener todas las palabras clave objetivo relativamente equilibradas, se calculó el número total de muestras que ésta debía contener y se seleccionó un número de muestras equitativo por cada palabra perteneciente a las palabras o números desconocidos de forma aleatoria. En

el siguiente enlace gráficos circulares[1], se encuentra un documento donde se pueden observar gráficos circulares para cada uno de los conjuntos de datos de cada modelo.

3.3.3 Arquitecturas seleccionadas

En la actualidad, con el éxito del aprendizaje profundo en una variedad de tareas de reconocimiento, los enfoques basados en redes neuronales se han vuelto populares para mejorar los métodos de reconocimiento de palabras clave al obtener modelos con mejor rendimiento, bajo consumo de memoria y costo computacional [93, 94]. Especialmente, en muchas investigaciones recientes han sugerido el uso de redes neuronales convolucionales (CNN) [93, 94, 99, 101, 102] y redes neuronales recurrentes RNN [93, 94, 98]; estas últimas también se han combinado con capas convolucionales para conformar las redes neuronales convolucionales recurrentes (CRNN) [92, 93, 94]. En la Tabla. 3.13 se observan las tasas de reconocimiento para las arquitecturas mencionadas anteriormente, las cuales fueron obtenidas en 2 investigaciones donde se realiza la comparación de rendimiento entre éstas utilizando el corpus de audio Speech Commands [106].

Arquitectura	**% Exactitud**	
	[93]	[94]
CNN	92.7%	96.0%
GRU	93.7%	97.2%
CRNN	95.0%	97.5%

Tabla 3.13: *Tasas de reconocimiento para diferentes arquitecturas de redes neuronales profundas*

En aras de realizar una comparativa del desempeño entre las diferentes arquitecturas, exploramos 3 de ellas para entrenar a nuestros modelos: CNN (ver Fig. 3.9); RNN con celdas GRU (ver Fig. 3.10); CRNN con celdas GRU (ver Fig. 3.11). Los hiperparámetros para cada arquitectura de red neuronal de nuestros modelos fueron tomados de la investigación [94], la cual presenta las mejores tasas de reconocimiento para cada arquitectura seleccionada.

[1] https://bit.ly/2Z5lCUb

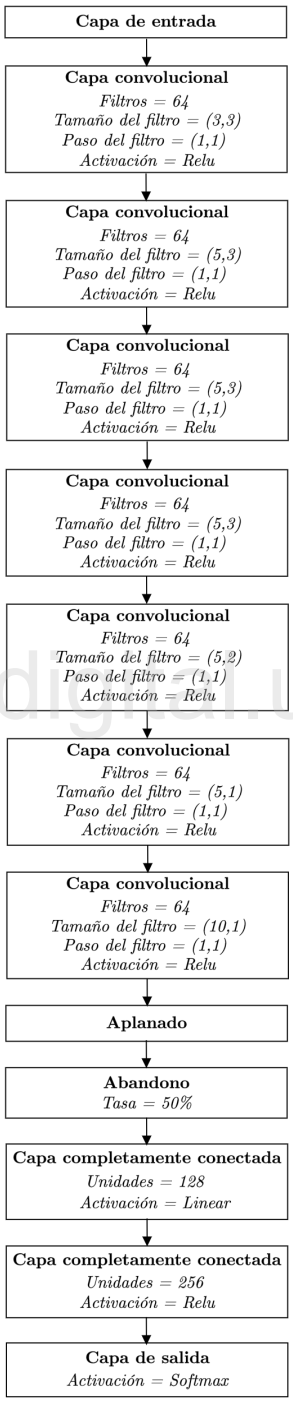


Figura 3.9: *Arquitectura de red convolucional utilizada [94]*

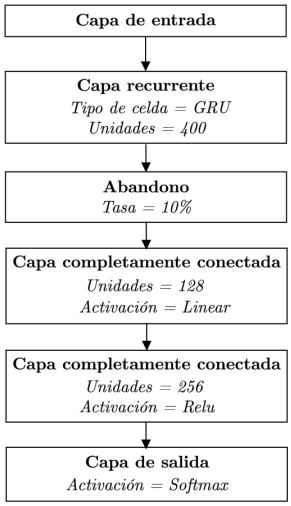


Figura 3.10: *Arquitectura de red recurrente utilizada [94]*

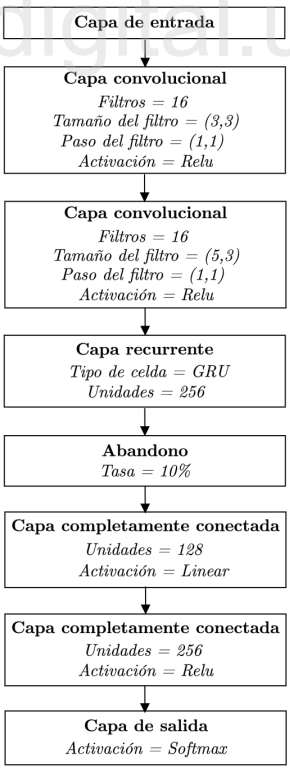


Figura 3.11: *Arquitectura de red convolucional recurrente utilizada [94]*

3.3.4 Preprocesamiento del audio

Para hacer frente a la falta de disponibilidad de una gran cantidad de datos del habla de los niños, hemos explorado la viabilidad de crear artificialmente más datos que sean acústicamente iguales o similares al habla de los niños mediante el aumento de datos. El aumento de datos es una técnica popular para aumentar el tamaño de los conjuntos de entrenamiento mediante la aplicación de transformaciones a las muestras originales para crear nuevas muestras que mantienen la etiqueta de la muestra original. El aumento de datos en el reconocimiento del habla es un método eficaz para reducir el desajuste, mejorar la solidez de los modelos y evitar el ajuste excesivo [109].

En este proyecto de grado, exploramos la perturbación de la velocidad y la inyección de ruido de fondo. Con la intención de estudiar el efecto de aumento de datos en el rendimiento de los modelos hemos probado 4 casos:

1. Conjunto entrenamiento sin aumento + conjunto de prueba sin ruido
2. Conjunto de entrenamiento aumentado + conjunto de prueba sin ruido
3. Conjunto entrenamiento sin aumento + conjunto de prueba con ruido
4. Conjunto de entrenamiento aumentado + conjunto de prueba con ruido

Para la perturbación de la velocidad, generamos 6 muestras de audios de una muestra de audio original perteneciente al conjunto de entrenamiento, cambiando la velocidad de las muestras de audio por un factor α seleccionado uniformemente al azar en el rango [0.85, 1.15]. Escogimos esta técnica de aumento de datos basados en el hecho de que, en el caso de los niños, éstos exhiben una mayor variabilidad en la velocidad del habla [110]. Los sistemas de reconocimiento del habla suelen funcionar bien en condiciones de voz limpia. Sin embargo, su rendimiento se degrada significativamente en condiciones ruidosas. Para mejorar la solidez de los modelos al ruido, corrompimos artificialmente cada muestra de audio con ruidos de fondo que pueden encontrarse en una primaria, como por ejemplo, ruido dentro de un salón de clases, ruido en los pasillos de una primaria o ruido en el patio de recreo. Cada muestra de audio fue combinada con ruido de fondo con una relación señal/ruido escogida uniformemente al azar entre el rango de [5, 30] dB tanto para el conjunto de entrenamiento como para el conjunto

de prueba. Para este último, las muestras de audio de ruido de fondo eran distintas a las usadas con el conjunto de datos de entrenamiento para evitar la fuga de datos.

Las muestras de audio en el corpus LaSDAICVI no poseen la misma duración; esto es un problema para arquitecturas convolucionales donde la capa entrada debe recibir entradas con un tamaño estándar. Por lo tanto, para que todas las muestras de audio de los diferentes conjuntos de datos posean la misma duración, decidimos estandarizar la duración en 2 segundos, tiempo suficiente para garantizar que cualquier muestra de audio del corpus se escuche completamente y no se trunque. Una forma de lograr que todas las muestras de audio posean la misma duración es rellenar con ceros (silencio) al final de la muestra de audio hasta alcanzar la duración deseada. En nuestro caso, debido a que cada muestra de audio en el corpus LaSDAICVI solo representa la palabra o número pronunciado por el niño, al aplicar la técnica de relleno con ceros al final de cada muestra, obtendríamos muestras de audio con la mayor cantidad de información al inicio de cada muestra. Por lo tanto, decidimos crear una muestra de audio silenciosa de 2 segundos, a la cual le insertamos la muestra de audio original en una posición aleatoria. De esta forma se debería ayudar a los modelos a aprender una representación más invariante en el tiempo de las palabras clave, ya que pueden aparecer en cualquier lugar dentro de la muestra de 2 segundos. Una vez estandarizadas todas las muestras de audio, se agregó un porcentaje de muestras de audio silenciosas con la etiqueta _silencio_ en cada uno de los conjuntos de datos.

3.3.5 Extracción de características

Los coeficientes cepstrales de frecuencia Mel (MFCC) están entre las características que más se utilizan comúnmente en el reconocimiento de voz basado en aprendizaje profundo, que se adapta de las técnicas tradicionales de procesamiento de voz. La extracción de características utilizando MFCC implica traducir la señal de voz en el dominio del tiempo en un conjunto de coeficientes espectrales en el dominio de la frecuencia, lo que permite la compresión de dimensionalidad de la señal de entrada. Para extraer MFCC, son necesarios los siguientes pasos [111]:

1. La transformada discreta de Fourier (DFT, por sus siglas en inglés) es calculada. Esta es usada para derivar la representación de la señal en el dominio de la frecuencia (espectral), la cual sirve como entrada para la obtención de muchas características importantes.

   Dada una señal discreta en el dominio del tiempo $x(n), n = 0, ..., N-1$, con N muestras de longitud, su DFT es calculada como sigue:

$$X(k) = \sum_{n=0}^{N-1} x(n) exp(-j\frac{2\pi}{N}kn),\ k = 0, ..., N-1,\ donde\ j \equiv \sqrt{-1} \tag{3.1}$$

2. El espectro resultante es utilizado como entrada a un banco de filtros de la escala de Mel que consiste en L filtros. Los filtros usualmente tienen una frecuencia triangular superpuesta. La escala de Mel introduce una función de distorsión de frecuencia (ver Fig. 3.12) que intenta ajustarse a ciertas observaciones psicoacústicas. A través de los años varias funciones de distorsión de frecuencias han sido propuestas, por ejemplo:

$$f_w = 2595 * log(1 + f/700) \tag{3.2}$$

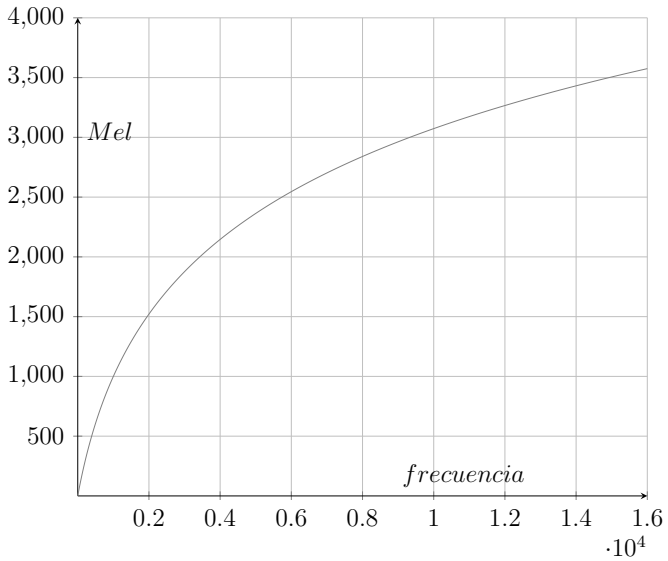


Figura 3.12: *Función de distorsión de frecuencias* f_w

Si $\tilde{O}_k, k = 1, ..., L$, es la potencia en la salida del k-ésimo filtro, entonces los MFCCs están dados por la siguiente ecuación

$$C_m = \sum_{k=1}^{L} (log\tilde{O}_k) cos[m(k - \frac{1}{2})\frac{\pi}{L}],\ m = 1, ..., L. \tag{3.3}$$

Para el paso de extracción de características, para todos los modelos, utilizamos 20 características de MFCC extraídas de una ventana de longitud de 1024 muestras (64 ms) con un paso de 512 muestras (32 ms) y 40 bancos de filtros, lo que da como resultado una matriz de características con 61 filas (marcos de tiempo) y 20 columnas (características MFCC) por cada muestra de audio de 2 segundos.

3.3.6 Implementación

Para la implementación de cada uno de nuestros modelos, hicimos uso de la biblioteca Keras de Tensorflow [112]. Para el entrenamiento seleccionamos un tamaño de lote de 128 muestras, una tasa de aprendizaje de 10^{-5} y la optimización estocástica de Adam [113]. Cada modelo fue entrenado hasta alcanzar la convergencia, por lo tanto, las épocas de entrenamiento varían de 300 a 400 épocas dependiendo de la arquitectura. Utilizamos el punto de control de la menor perdida de validación para guardar los modelos con el mejor rendimiento. En el siguiente enlace curvas de aprendizaje[2], se encuentra un documento donde se pueden observar las curvas de aprendizaje para cada uno de los modelos entrenados.

Debido a que los modelos reconocedores de palabras clave deben desplegarse en el dispositivo, es deseable que tales modelos tengan un consumo de memoria bajo, así como también un costo computacional bajo para que puedan implementarse en dispositivos de bajo consumo energético y rendimiento limitado. En un escenario de interacción humano-robot estas características son muy deseables, ya que por lo general, en estos escenarios se deben realizar diferentes tareas de reconocimiento, lo cual es un inconveniente para el limitado almacenamiento y procesamiento interno de los robots.

[2] https://bit.ly/3qW2svB

Por tal motivo, una vez entrenados los modelos, utilizamos el conjunto de herramientas de TensorFlow Lite para optimizarlos, y obtener modelos con un tamaño reducido, un menor consumo de energía y una velocidad de inferencia más rápida para que puedan ser ejecutados de forma eficiente en dispositivos con recursos de procesamiento y memoria limitados.

En la Fig. 3.13 se puede observar un diagrama de bloques donde se muestra el proceso de entrenamiento de un modelo. En los casos donde se deba mantener el conjunto de entrenamiento sin modificaciones, los bloques de aumentos de datos (perturbación de la velocidad y adición de ruido de fondo) son obviados.

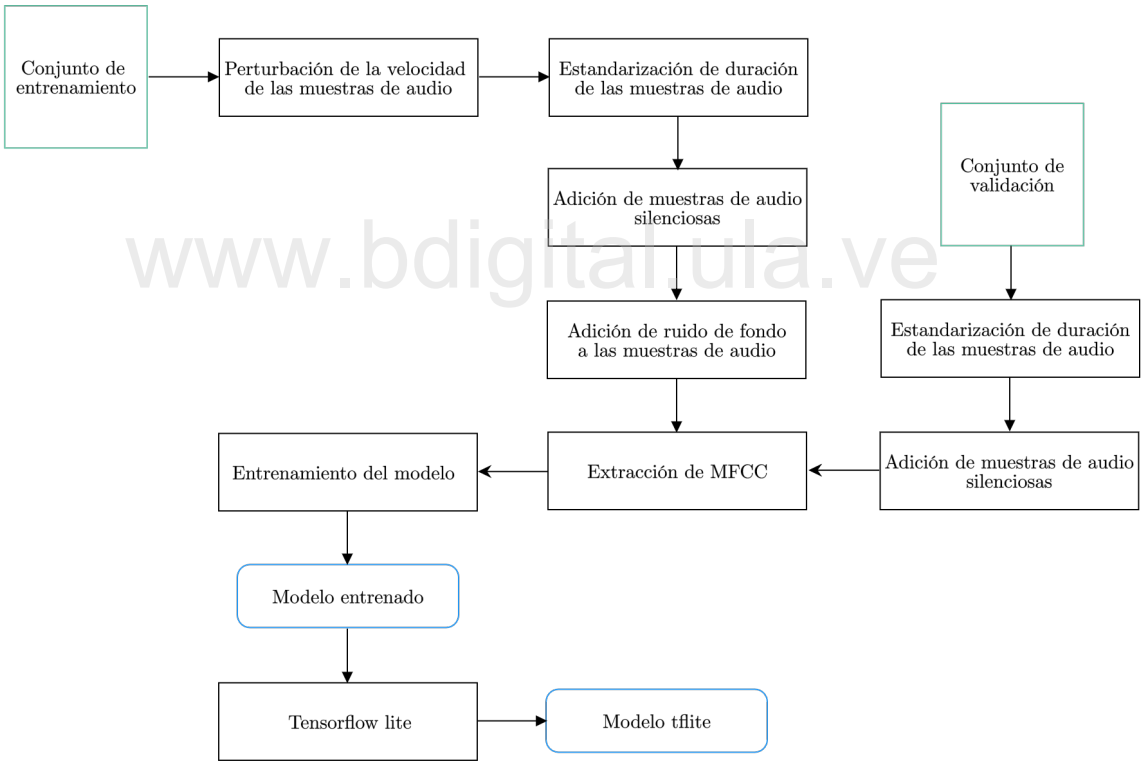


Figura 3.13: *Diagrama de bloques del proceso de entrenamiento*

Capítulo 4

Pruebas y análisis de los resultados

Una vez que un modelo de aprendizaje profundo es entrenado, una tarea frecuente es probar el modelo con datos distintos a los de entrenamiento, con la finalidad de medir su desempeño para predecir datos no antes vistos. Para evaluar qué tan bien se desempeñan nuestros modelo, hemos decidido utilizar la exactitud como métrica principal, es decir, la proporción de decisiones correctas sobre el número total de predicciones realizadas:

$$Exactitud = \frac{Núm.\ de\ predicciones\ correctas}{Núm.\ total\ de\ predicciones} \tag{4.1}$$

Asimismo, calculamos el área bajo la curva (AUC) de la curva de característica operativa del receptor (ROC), donde el eje x y el eje y denotan las tasas de falsa alarma (la probabilidad de dar un resultado positivo cuando el valor verdadero sea negativo) y falso rechazo (la probabilidad de dar un resultado negativo cuando el valor verdadero sea positivo), respectivamente. Una menor área bajo la curva (AUC) significa que el modelo perdería menos palabras clave objetivo en promedio para varias tasas de falsas alarmas, lo cual es fundamental para una buena experiencia de usuario en los sistemas de reconocimiento de palabras clave. Aunque las curvas ROC se utilizan normalmente para evaluar clasificadores binarios, ampliamos éste a la clasificación de clases múltiples mediante micro promedio sobre todas las clases por modelo, de forma similar a otros trabajos [99, 102].

De igual manera, calculamos varias métricas para realizar un análisis más profundo de cada modelo, y observar su comportamiento para cada clase individual. Para esto, obtenemos la matriz de confusión y calculamos la precisión, la sensibilidad y el puntaje F1. La precisión, permite estimar el costo de los falsos positivos en la clasificación; la sensibilidad, permite estimar el costo de los falsos negativos en la clasificación; por último, el puntaje F1, permite evaluar la exactitud en función de la precisión y de la sensibilidad, en otras palabras, evaluar la exactitud en función el costo de los falsos positivos y falsos negativos.

$$Precisión = \frac{Núm.\ de\ predicciones\ correctas\ que\ realmente\ son\ correctas}{Núm.\ total\ predicciones\ marcadas\ como\ correctas} \tag{4.2}$$

$$Sensibilidad = \frac{Núm.\ de\ predicciones\ correctas\ que\ realmente\ son\ correctas}{Núm.\ muestras\ correctas\ en\ el\ conjunto\ de\ prueba} \tag{4.3}$$

$$Puntaje\ F1 = 2 \times \frac{Precisión \times Sensibilidad}{Precisión + Sensibilidad} \tag{4.4}$$

En la Fig. 4.1 se puede observar un diagrama de bloques donde se muestra el proceso de evaluación de un modelo. En los casos donde se deba mantener el conjunto de datos de prueba sin modificaciones, el bloque de adición de ruido de fondo es obviado. Para ser consistentes con el proceso de evaluación, los modelos fueron evaluados utilizando los mismos conjuntos de prueba.

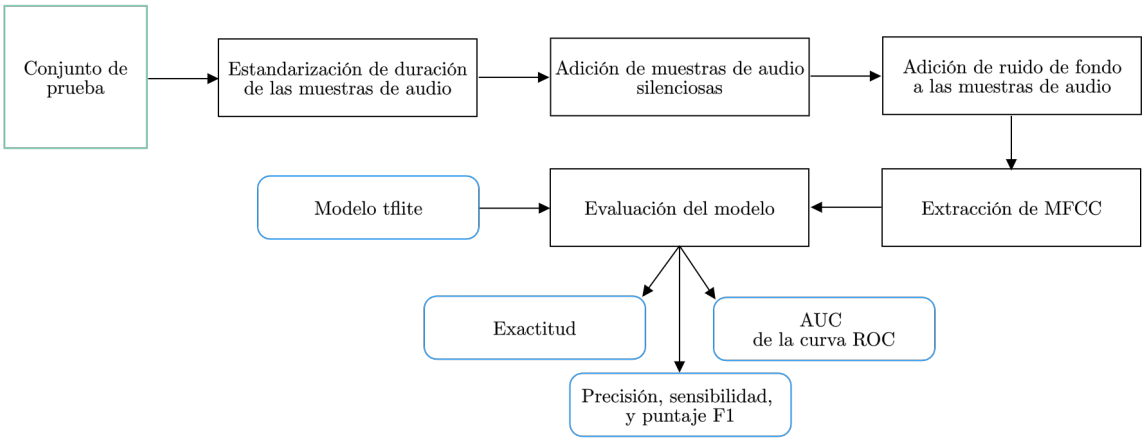


Figura 4.1: *Diagrama de bloques del proceso de evaluación*

4.1 Exactitud y micro promedio AUC

En la Tabla 4.1, se pueden observar los resultados obtenidos para exactitud y micro promedio AUC de cada uno de los modelos entrenados. En ella podemos observar que para los modelos de la *tabla del 2* y la *tabla del 8* con arquitectura CRNN entrenados con el conjunto de entrenamiento aumentado, se obtienen los mejores resultados para ambos conjuntos de prueba. Por otro lado, los modelos de la *tabla del 3*, la *tabla del 4*, la *tabla del 5*, *interacción* y *activación*, obtienen los mejores resultados para ambos conjuntos de prueba con la arquitectura GRU, cuando son entrenados con el conjunto de entrenamiento aumentado.

Aunque los modelos de la *tabla del 6* y la *tabla del 7* obtienen un valor de exactitud mayor para los modelos entrenados con el conjunto de entrenamiento aumentado con arquitecturas GRU comparado con los modelos con arquitecturas CRNN, podemos apreciar que el modelo con arquitectura CRNN obtiene un valor AUC menor para el conjunto de pruba sin ruido en comparación al modelo GRU. Estos resultados se deben a la forma en la que se calcula el micro promedio, en donde todas las muestras contribuyen por igual a la métrica promediada final. Por lo tanto, las clases con más muestras son las más dominantes, indicando así que en la prueba sin ruido los modelos con arquitectura CRNN obtienen tasas de falsos positivos y tasas de falsos negativos más bajas para la clases con el mayor número de muestras.

Para todos los modelos con el mejor rendimiento, se obtuvo una exactitud mayor al 90%, siendo el modelo de *activación* quien obtuvo el valor de exactitud más alto para ambas pruebas, 99.38% para el conjunto de prueba sin ruido y 97.82% para el conjunto de prueba con ruido. Por el contrario, el modelo de la *tabla del 9* fue quien obtuvo los resultados más bajos de exactitud, 96.12% para el conjunto de prueba sin ruido y 93.13% para el conjunto de prueba con ruido.

Por último, se puede observar que todos los modelos entrenados con el conjunto de entrenamiento sin aumento de datos, degradan su rendimiento cuando el conjunto de datos de prueba contiene ruido en las muestras. Por el contrario, todos los modelos aumentan su rendimiento en ambas pruebas cuando se entrenan con el conjunto de entrenamiento aumentado.

Modelo	Arquitectura	Conjunto de entrenamiento	Exactitud		AUC	
			limpio	Ruido	limpio	Ruido
Tabla del 2	CNN	Original	91.98%	33.13%	0.0040756	0.2898468
		Aumentado	97.03%	92.53%	0.0006737	0.0032106
	GRU	Original	95.44%	41.42%	0.0023621	0.2288262
		Aumentado	97.10%	95.57%	0.0005807	0.0016160
	CRNN	Original	94.95%	32.02%	0.0024613	0.3147178
		Aumentado	**97.86%**	**96.33%**	**0.0004205**	**0.0011396**
Tabla del 3	CNN	Original	86.70%	34.13%	0.0070049	0.3216751
		Aumentado	94.52%	88.62%	0.0014326	0.0062828
	GRU	Original	93.83%	40.71%	0.0030460	0.2146977
		Aumentado	**97.19%**	**94.52%**	**0.0009039**	**0.0020528**
	CRNN	Original	94.17%	30.50%	0.0023101	0.3287901
		Aumentado	96.85%	93.76%	0.0011819	0.0026557
Tabla del 4	CNN	Original	88.40%	34.24%	0.0059607	0.2705737
		Aumentado	94.86%	89.93%	0.0013445	0.0054297
	GRU	Original	93.40%	43.47%	0.0029801	0.1931431
		Aumentado	**97.22%**	**95.00%**	**0.0007343**	**0.0028439**
	CRNN	Original	94.65%	34.65%	0.0012615	0.2589904
		Aumentado	96.46%	94.10%	0.0007367	0.0033074
Tabla del 5	CNN	Original	87.34%	33.15%	0.0064726	0.3296081
		Aumentado	93.29%	86.23%	0.0020826	0.0090982
	GRU	Original	92.53%	41.59%	0.0031124	0.2151512
		Aumentado	**97.79%**	**95.02%**	**0.0006384**	**0.0031350**
	CRNN	Original	91.56%	29.76%	0.0046750	0.3295046
		Aumentado	95.57%	91.76%	0.0012124	0.0038429
Tabla del 6	CNN	Original	89.48%	34.94%	0.0051206	0.2958438
		Aumentado	94.23%	89.28%	0.0015286	0.0046577
	GRU	Original	93.08%	37.92%	0.0029401	0.2073710
		Aumentado	**97.90%**	**95.12%**	0.0010652	**0.0019798**
	CRNN	Original	90.98%	31.89%	0.0033517	0.3176272
		Aumentado	96.13%	92.94%	**0.0007149**	0.0024674
Tabla del 7	CNN	Original	88.33%	48.66%	0.0065337	0.2035843
		Aumentado	93.21%	88.33%	0.0021198	0.0058384
	GRU	Original	94.58%	40.01%	0.0016338	0.1986852
		Aumentado	**97.25%**	**94.65%**	0.0006038	**0.0023521**
	CRNN	Original	94.17%	35.62%	0.0016905	0.2253169
		Aumentado	96.84%	94.23%	**0.0005875**	0.0030476
Tabla del 8	CNN	Original	88.57%	42.31%	0.0067912	0.2369026
		Aumentado	94.42%	89.39%	0.0014271	0.0050700
	GRU	Original	93.06%	40.82%	0.0023950	0.1975906
		Aumentado	95.03%	91.97%	0.0015942	0.0046266
	CRNN	Original	93.61%	36.39%	0.0022072	0.2485352
		Aumentado	**96.87%**	**94.08%**	**0.0004184**	**0.0033212**
Tabla del 9	CNN	Original	87.76%	44.08%	0.0076615	0.2272587
		Aumentado	93.33%	88.10%	0.0021172	0.0061336
	GRU	Original	93.13%	44.63%	0.0022186	0.1808663
		Aumentado	**96.12%**	**93.13%**	**0.0007106**	**0.0025360**
	CRNN	Original	91.63%	42.99%	0.0034020	0.1968283
		Aumentado	95.37%	91.97%	0.0008750	0.0046881
Activación	CNN	Original	96.26%	44.86%	0.0039547	0.2413360
		Aumentado	97.51%	93.77%	0.0006793	0.0045710
	GRU	Original	97.51%	61.68%	0.0006163	0.2140604
		Aumentado	**99.38%**	**97.82%**	**0.0000970**	**0.0007085**
	CRNN	Original	98.13%	45.17%	0.0010772	0.4432338
		Aumentado	97.82%	94.70%	0.0006648	0.0071768
Interacción	CNN	Original	89.62%	40.35%	0.0066824	0.2486499
		Aumentado	95.54%	88.85%	0.0007994	0.0055911
	GRU	Original	94.63%	43.07%	0.0031247	0.2023213
		Aumentado	**98.33%**	**96.86%**	**0.0003775**	**0.0008548**
	CRNN	Original	96.10%	35.89%	0.0006416	0.2688441
		Aumentado	97.63%	95.12%	0.0004056	0.0010628

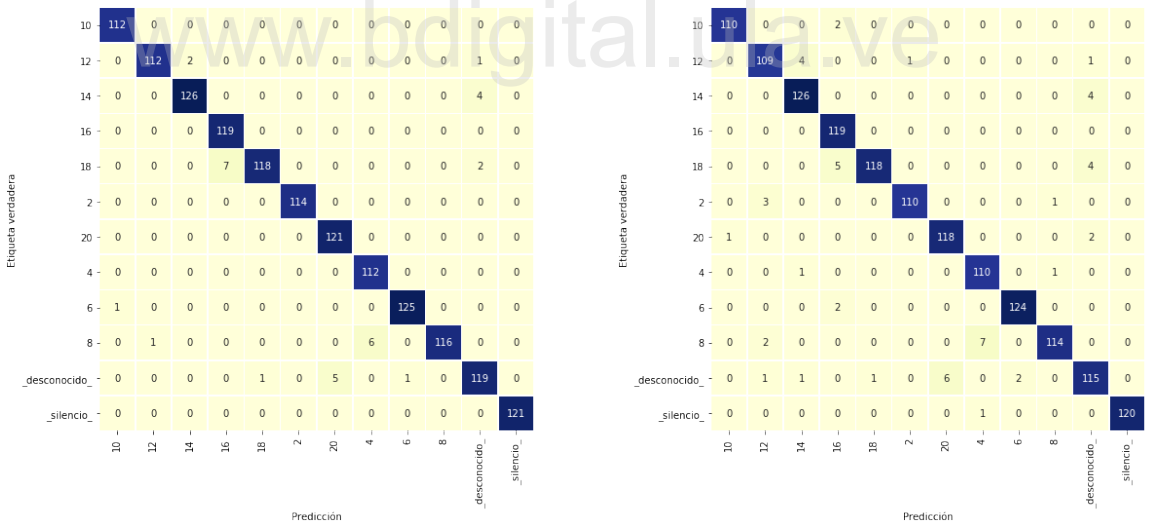
Tabla 4.1: *Exactitud y micro promedio AUC de los modelos de reconocimiento de palabras clave*

4.2 Métricas de precisión, sensibilidad y puntaje F1

A continuación, se presentan las matrices de confusión obtenidas para los modelos con mejor rendimiento de la Tabla 4.1, junto con las métricas: precisión, sensibilidad y puntaje F1 para cada uno de los conjuntos de prueba.

4.2.1 Modelo de la tabla del 2

En la Fig. 4.2, se pueden observar las matrices de confusión para el modelo de la *tabla del 2* con arquitectura CRNN entrenado con el conjunto de entrenamiento aumentado cuando se prueba con el conjunto de prueba sin ruido (Fig. 4.2a) y con ruido (Fig. 4.2b). Podemos notar que para ambas pruebas, el modelo tiende a clasificar incorrectamente en mayor medida el número 18 con el número 16, el número 8 con el número 4, y la clase _desconocido_ con el número 20.



(a) *Conjunto de prueba sin ruido* **(b)** *Conjunto de prueba con ruido*

Figura 4.2: *Matrices de confusión para el modelo de la tabla del 2 con arquitectura CRNN entrenado con el conjunto de entrenamiento aumentado*

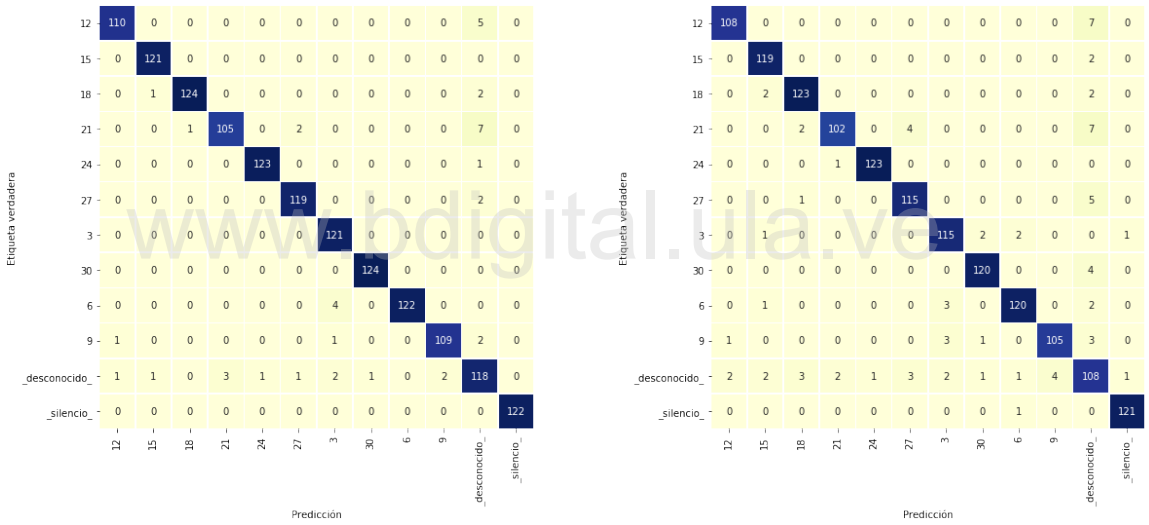
En la Tabla 4.2 podemos apreciar que cuando se prueba con el conjunto de prueba sin ruido, los valores más bajos de precisión y puntaje F1 corresponden a la clase _desconocido_, mientras que el número 18 presenta el valor más bajo en sensibilidad. Lo que indica que el modelo tiende a clasificar incorrectamente otras clases como _desconocido_ y rechaza el número 18 clasificándolo como otra clase distinta. Para el conjunto de prueba con ruido, la clase _desconocido_ obtiene los valores más bajos en las 3 distintas métricas, lo que indica que el modelo presenta dificultades al clasificar números que no se encuentren dentro de los resultados de la tabla de multiplicación del 2 en condiciones ruidosas.

Etiqueta	**Conjunto de prueba sin ruido**			**Conjunto de prueba con ruido**		
	Precisión	**Sensibilidad**	**Puntaje F1**	**Precisión**	**Sensibilidad**	**Puntaje F1**
2	100.00%	100.00%	100.00%	99.10%	96.49%	97.78%
4	94.92%	100.00%	97.39%	93.22%	98.21%	95.65%
6	99.21%	99.21%	99.21%	98.41%	98.41%	98.41%
8	100.00%	94.31%	97.07%	98.28%	92.68%	95.40%
10	99.12%	100.00%	99.56%	99.10%	98.21%	98.65%
12	99.12%	97.39%	98.25%	94.78%	94.78%	94.78%
14	98.44%	96.92%	97.67%	95.45%	96.92%	96.18%
16	94.44%	100.00%	97.14%	92.97%	100.00%	96.36%
18	99.16%	**92.91**%	95.93%	99.16%	92.91%	95.93%
20	96.03%	100.00%	97.98%	95.16%	97.52%	96.33%
desconocido	**94.44**%	94.44%	**94.44**%	**91.27**%	**91.27**%	**91.27**%
silencio	100.00%	100.00%	100.00%	100.00%	99.17%	99.59%

Tabla 4.2: *Precisión, sensibilidad y puntaje F1 para el modelo de la tabla del 2 con arquitectura CRNN entrenado con el conjunto de entrenamiento aumentado*

4.2.2 Modelo de la tabla del 3

En la Fig. 4.3, se pueden observar las matrices de confusión para el modelo de la *tabla del 3* con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado cuando se prueba con el conjunto de prueba sin ruido (Fig. 4.3a) y con ruido (Fig. 4.3b). Podemos notar que para el conjunto de prueba sin ruido, el modelo tiende a clasificar incorrectamente en mayor medida el número 12 y el número 21 con la clase _desconocido_. Para el conjunto de prueba con ruido, se mantienen los mismos errores de clasificación, pero además, el modelo clasifica incorrectamente el número 27 con la clase _desconocido_.



(a) *Conjunto de prueba sin ruido* **(b)** *Conjunto de prueba con ruido*

Figura 4.3: *Matrices de confusión para el modelo de la tabla del 3 con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado*

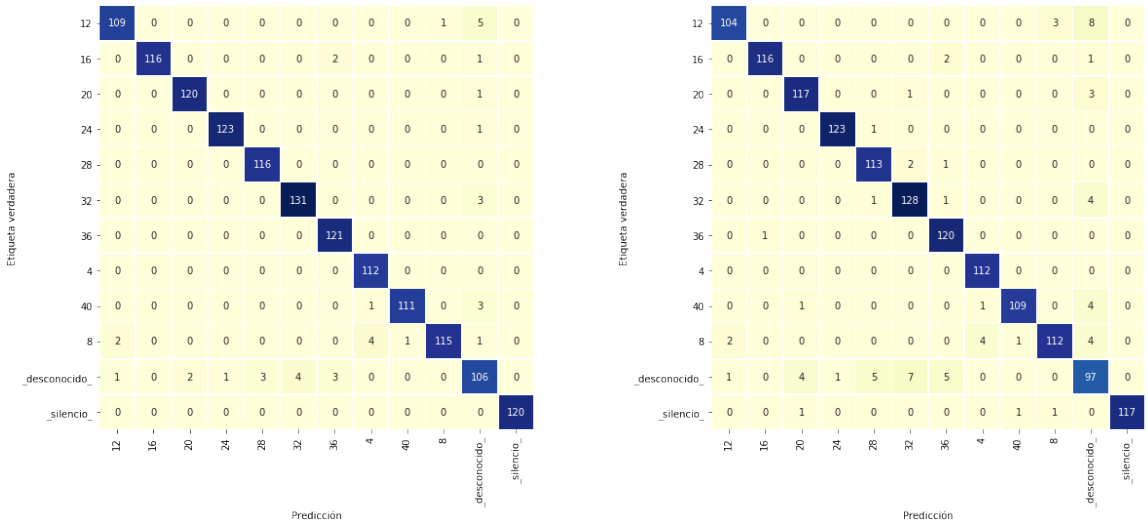
Etiqueta	Conjunto de prueba sin ruido			Conjunto de prueba con ruido		
	Precisión	**Sensibilidad**	**Puntaje F1**	**Precisión**	**Sensibilidad**	**Puntaje F1**
3	94.53%	100.00%	97.19%	93.50%	96.04%	94.26%
6	100.00%	96.83%	98.39%	96.77%	95.24%	96.00%
9	98.20%	96.46%	97.32%	96.33%	92.92%	94.59%
12	98.21%	95.65%	96.92%	97.30%	93.91%	95.58%
15	98.37%	100.00%	99.18%	95.20%	98.35%	96.75%
18	99.20%	97.64%	98.41%	95.35%	96.85%	96.09%
21	97.22%	91.30%	94.17%	97.14%	88.70%	92.73%
24	99.19%	99.19%	99.19%	99.19%	99.19%	99.19%
27	97.54%	98.35%	97.94%	94.26%	95.04%	94.65%
30	96.03%	100.00%	97.98%	96.77%	96.77%	96.77%
desconocido	**86.13**%	**90.77**%	**88.39**%	**77.14**%	**83.08**%	**80.00**%
silencio	100.00%	100.00%	100.00%	98.37%	99.18%	98.78%

Tabla 4.3: *Precisión, sensibilidad y puntaje F1 para el modelo de la tabla del 3 con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado*

En la Tabla 4.3 podemos apreciar que para ambos conjuntos de prueba, los valores más bajos de precisión, sensibilidad y puntaje F1 corresponden a la clase _desconocido_. lo que indica que el modelo de la *tabla del 3* presenta dificultades al clasificar números que no se encuentren dentro de los resultados de la tabla de multiplicación del 3.

4.2.3 Modelo de la tabla del 4

En la Fig. 4.4, se pueden observar las matrices de confusión para el modelo de la *tabla del 4* con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado cuando se prueba con el conjunto de prueba sin ruido (Fig. 4.4a) y con ruido (Fig. 4.4b). Podemos notar que para ambas pruebas, el modelo tiende a clasificar incorrectamente en mayor medida el número 12 con la clase _desconocido_, y la clase _desconocido_ con los números 28, 32, 36.



(a) *Conjunto de prueba sin ruido*

(b) *Conjunto de prueba con ruido*

Figura 4.4: *Matrices de confusión para el modelo de la tabla del 4 con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado*

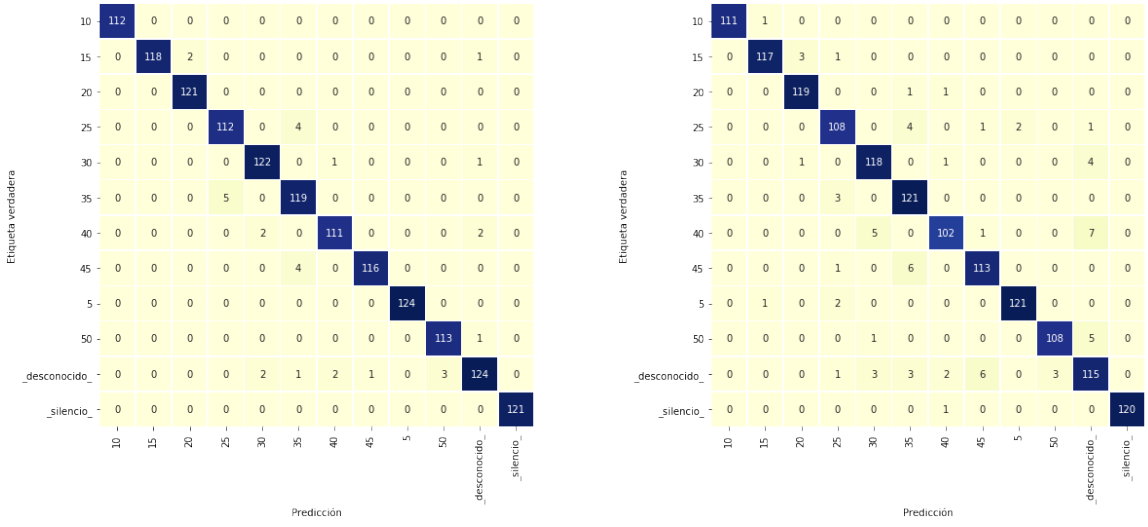
Etiqueta	Conjunto de prueba sin ruido			Conjunto de prueba con ruido		
	Precisión	**Sensibilidad**	**Puntaje F1**	**Precisión**	**Sensibilidad**	**Puntaje F1**
4	95.73%	100.00%	97.82%	95.73%	100.00%	97.82%
8	99.14%	93.50%	96.23%	96.55%	91.06%	93.72%
12	97.32%	94.78%	96.04%	97.20%	90.43%	93.69%
16	100.00%	97.48%	98.72%	99.15%	97.48%	98.31%
20	98.36%	99.17%	98.77%	95.12%	96.69%	95.90%
24	99.19%	99.19%	99.19%	99.19%	99.19%	99.19%
28	97.48%	100.00%	98.72%	94.17%	97.41%	95.76%
32	97.04%	97.76%	97.40%	92.75%	95.52%	94.12%
36	96.03%	100.00%	97.98%	93.02%	99.17%	96.00%
40	99.11%	96.52%	97.80%	98.20%	94.78%	96.46%
desconocido	**87.60%**	**88.33%**	**87.97%**	**80.17%**	**80.83%**	**80.50%**
silencio	100.00%	100.00%	100.00%	100.00%	97.50%	98.73%

Tabla 4.4: *Precisión, sensibilidad y puntaje F1 para el modelo de la tabla del 4 con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado*

En la Tabla 4.4 podemos apreciar que para ambos conjuntos de prueba, los valores más bajos de precisión, sensibilidad y puntaje F1 corresponden a la clase _desconocido_, lo que indica que el modelo de la *tabla del 4* presenta dificultades al clasificar números que no se encuentren dentro de los resultados de la tabla de multiplicación del 4.

4.2.4 Modelo de la tabla del 5

En la Fig. 4.5, se pueden observar las matrices de confusión para el modelo de la *tabla del 5* con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado cuando se prueba con el conjunto de prueba sin ruido (Fig. 4.5a) y con ruido (Fig. 4.5b). Podemos notar que para el conjunto de prueba sin ruido, el modelo tiende a confundir mayormente el número 35 con el número 25, y el número 45 con el número 35. Para el conjunto de prueba con ruido, el modelo confunde nuevamente el número 35 con el número 25, y el número 45 con el número 35, pero además, aumenta los errores de clasificación al confundir los números 40 y 50 con la clase _desconocido_, y esta última con el número 45.



(a) *Conjunto de prueba sin ruido* **(b)** *Conjunto de prueba con ruido*

Figura 4.5: *Matrices de confusión para el modelo de la tabla del 5 con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado*

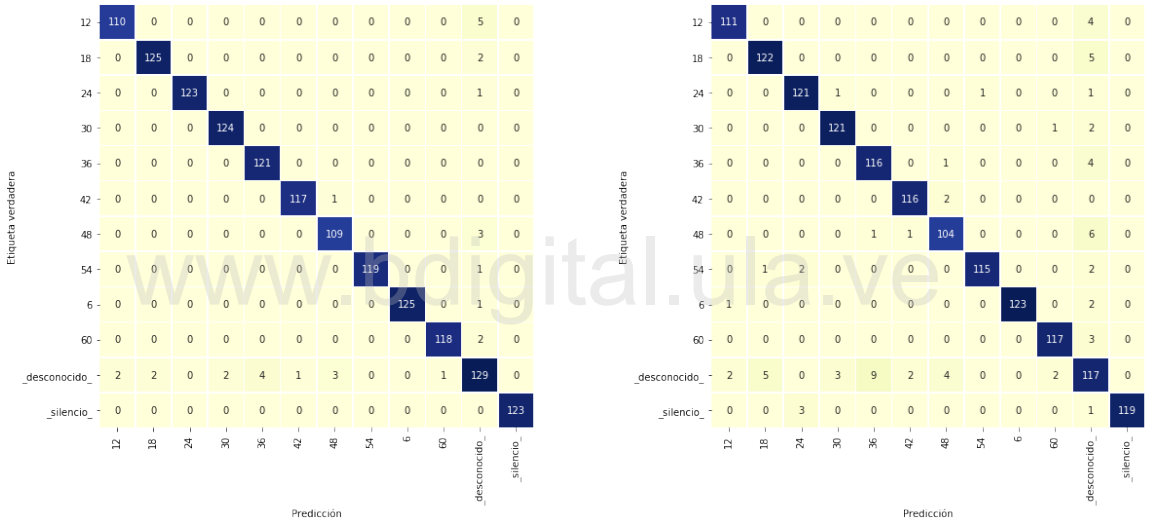
En la Tabla 4.5 podemos apreciar que para el conjunto de prueba sin ruido el número 35 obtiene los valores más bajos para precisión y puntaje F1, mientras que la clase _desconocido_ obtiene el valor más bajo de sensibilidad. Lo que indica que el modelo tiende a clasificar otras clases como el número 35 y rechazar la clase _desconocido_ clasificándola como otra clase. En el conjunto de prueba con ruido los valores más bajos de precisión, sensibilidad y puntaje F1 corresponden a la clase _desconocido_, lo que indica que para este caso, el modelo presenta dificultades al clasificar números que no se encuentren dentro de los resultados de la tabla de multiplicación del 5 en condiciones ruidosas.

Etiqueta	**Conjunto de prueba sin ruido**			**Conjunto de prueba con ruido**		
	Precisión	**Sensibilidad**	**Puntaje F1**	**Precisión**	**Sensibilidad**	**Puntaje F1**
5	100.00%	100.00%	100.00%	98.37%	97.58%	97.98%
10	100.00%	100.00%	100.00%	100.00%	99.11%	99.55%
15	100.00%	97.52%	98.74%	98.32%	96.69%	97.50%
20	98.37%	100.00%	99.18%	96.75%	98.35%	97.54%
25	95.73%	96.55%	96.14%	93.10%	93.10%	93.10%
30	96.83%	98.39%	97.60%	92.91%	95.16%	94.02%
35	**92.97**%	95.97%	**94.44**%	89.63%	97.58%	93.44%
40	97.37%	96.52%	96.94%	95.33%	88.70%	91.98%
45	99.15%	96.67%	97.89%	93.39%	94.17%	93.78%
50	97.41%	99.12%	98.26%	97.30%	94.74%	96.00%
desconocido	96.12%	**93.23**%	94.66%	**87.12**%	**86.47**%	**86.79**%
silencio	100.00%	100.00%	100.00%	100.00%	99.17%	99.59%

Tabla 4.5: *Precisión, sensibilidad y puntaje F1 para el modelo de la tabla del 5 con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado*

4.2.5 Modelo de la tabla del 6

En la Fig. 4.6, se pueden observar las matrices de confusión para el modelo de la *tabla del 6* con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado cuando se prueba con el conjunto sin ruido (Fig. 4.6a) y con ruido (Fig. 4.6b). Se puede apreciar que para ambos conjuntos de prueba, los errores de clasificación se presentan al confundir números con la clase _desconocido_ y viceversa. En el conjunto de prueba, se confunden en mayor medida los números 12, 18 y 48 con la clase _desconocido_, y esta última es confundida mayormente con el número 36.



(a) *Conjunto de prueba sin ruido* **(b)** *Conjunto de prueba con ruido*

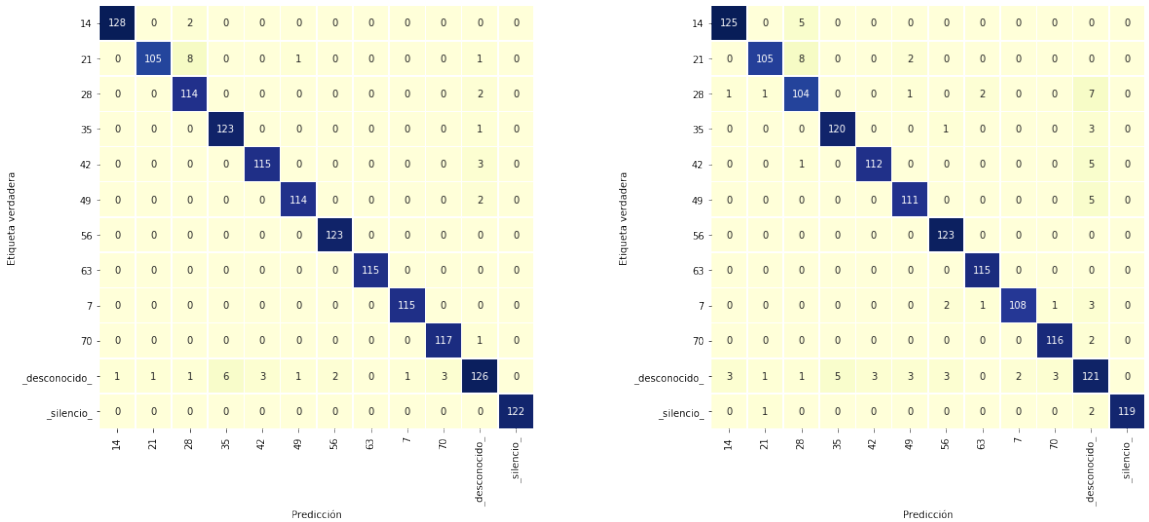
Figura 4.6: *Matrices de confusión para el modelo de la tabla del 6 con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado*

En la Tabla 4.6 se muestra que para ambos conjuntos de prueba, los valores más bajos en las métricas corresponden a la clase _desconocido_ , lo que indica que el modelo de la *tabla del 6* presenta dificultades al clasificar números que no se encuentren dentro de los resultados de la tabla de multiplicación del 6.

Etiqueta	Conjunto de prueba sin ruido			Conjunto de prueba con ruido		
	Precisión	Sensibilidad	Puntaje F1	Precisión	Sensibilidad	Puntaje F1
6	100.00%	99.21%	99.60%	100.00%	97.62%	98.80%
12	98.21%	95.65%	96.92%	97.32%	96.52%	96.94%
18	98.43%	98.43%	98.43%	95.31%	96.06%	95.69%
24	100.00%	99.19%	99.60%	96.03%	97.58%	96.80%
30	98.41%	100.00%	99.20%	96.80%	97.58%	97.19%
36	96.80%	100.00%	98.37%	92.06%	95.87%	93.93%
42	99.15%	99.15%	99.15%	97.48%	98.31%	97.89%
48	96.46%	97.32%	96.89%	93.69%	92.86%	93.27%
54	100.00%	99.17%	99.58%	99.14%	95.83%	97.46%
60	99.16%	98.33%	98.74%	97.50%	97.50%	97.50%
desconocido	**89.58%**	**89.58%**	**89.58%**	**79.59%**	**81.25%**	**80.41%**
silencio	100.00%	100.00%	100.00%	100.00%	96.75%	98.35%

Tabla 4.6: *Precisión, sensibilidad y puntaje F1 para el modelo de la tabla del 6 con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado*

4.2.6 Modelo de la tabla del 7



(a) *Conjunto de prueba sin ruido* **(b)** *Conjunto de prueba con ruido*

Figura 4.7: *Matrices de confusión para el modelo de la tabla del 7 con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado*

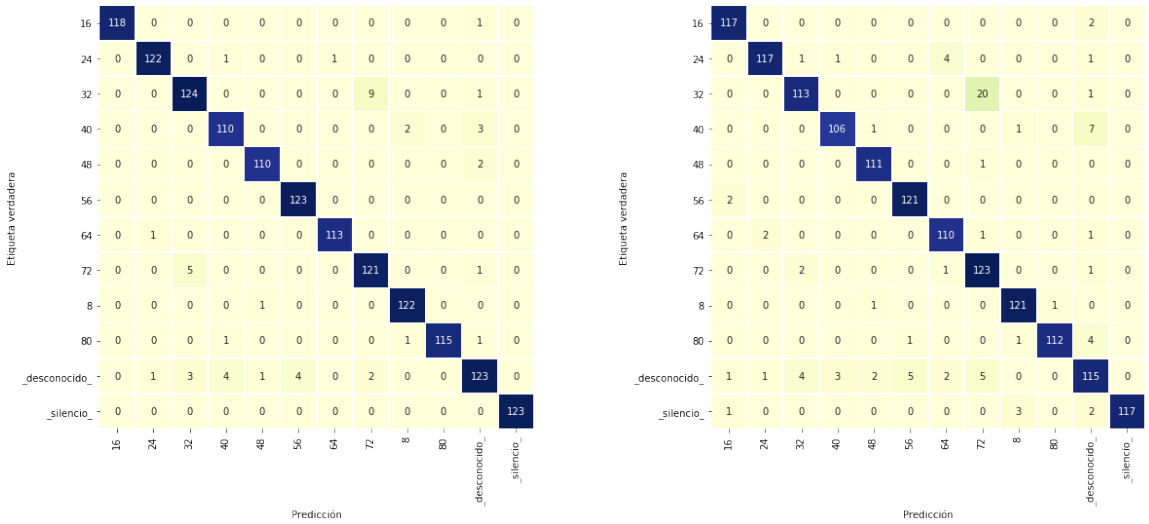
En la Fig. 4.7, se pueden observar las matrices de confusión para el modelo de la tabla del 7 con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado cuando se prueba con el conjunto sin ruido (Fig. 4.7a) y con ruido (Fig. 4.7b). Se puede apreciar que para el conjunto de prueba sin ruido, el modelo tiende a confundir en mayor medida el número 21 con el número 28, y la clase _desconocido_ con el número 35. En el conjunto de prueba con ruido, el modelo tiende a confundir nuevamente el número 21 con el número 28, y la clase _desconocido_ con el número 35. Pero además, aumentan los errores de clasificación de los números 28, 42, y 49 con la clase _desconocido_.

Etiqueta	**Conjunto de prueba sin ruido**			**Conjunto de prueba con ruido**		
	Precisión	**Sensibilidad**	**Puntaje F1**	**Precisión**	**Sensibilidad**	**Puntaje F1**
7	99.14%	100.00%	99.57%	98.18%	93.91%	96.00%
14	99.22%	98.46%	98.84%	96.90%	96.15%	96.53%
21	99.06%	91.30%	95.02%	97.22%	91.30%	94.17%
28	**91.20**%	98.28%	94.61%	87.39%	89.66%	88.51%
35	95.35%	99.19%	97.23%	96.00%	96.77%	96.39%
42	97.46%	97.46%	97.46%	97.39%	94.92%	96.14%
49	98.28%	98.28%	98.28%	94.87%	95.69%	95.28%
56	98.40%	100.00%	99.19%	95.35%	100.00%	97.62%
63	100.00%	100.00%	100.00%	97.46%	100.00%	98.71%
70	97.50%	99.15%	98.32%	96.67%	98.31%	97.48%
desconocido	92.65%	**86.90**%	**89.68**%	**81.76**%	**83.45**%	**82.59**%
silencio	100.00%	100.00%	100.00%	100.00%	97.54%	98.76%

Tabla 4.7: *Precisión, sensibilidad y puntaje F1 para el modelo de la tabla del 7 con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado*

En la Tabla 4.7 se muestra que para el conjunto de prueba sin ruido el valor más bajo de precisión corresponde al número 28, mientras que la clase _desconocido_ presenta los valores más bajos de sensibilidad y puntaje F1. Lo que indica que el modelo tiende a clasificar otras clases como el número 28 y rechazar la clase _desconocido_ clasificándola como otra clase. Para el conjunto de prueba con ruido, los valores más bajos en las 3 métricas corresponden a la clase _desconocido_, lo que indica que el modelo de la *tabla del 7* presenta dificultades al clasificar números que no se encuentren dentro de los resultados de la tabla de multiplicación del 7 en condiciones de ruidosas.

4.2.7 Modelo de la tabla del 8



(a) *Conjunto de prueba sin ruido* **(b)** *Conjunto de prueba con ruido*

Figura 4.8: *Matrices de confusión para el modelo de la tabla del 8 con arquitectura CRNN entrenado con el conjunto de entrenamiento aumentado*

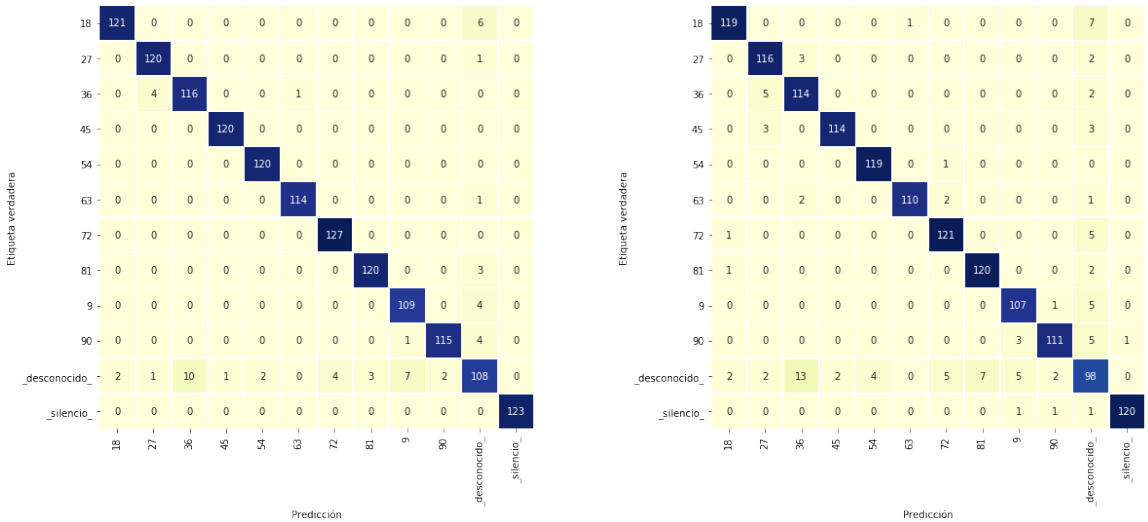
En la Fig. 4.8, se pueden observar las matrices de confusión para el modelo de la tabla del 8 con arquitectura CRNN entrenado con el conjunto de entrenamiento aumentado cuando se prueba con el conjunto sin ruido (Fig. 4.8a) y con ruido (Fig. 4.8b). Se puede apreciar que para ambas pruebas, el modelo confunde en mayor medida el número 32 con el número 72, siendo más evidente este error de clasificación en la prueba con ruido. Además, para la prueba con ruido, se aprecia un incremento de los errores de clasificación del número 40 con la clase _desconocido_.

En la Tabla 4.8 se muestra que para ambos conjuntos de prueba los valores más bajos de precisión corresponden al número 72, mientras que los valores de sensibilidad y puntaje F1 corresponden a la clase _desconocido_. Lo que indica que el modelo tiende a clasificar otras clases como el número 72 y rechazar la clase _desconocido_ clasificándola como otra clase.

Etiqueta	Conjunto de prueba sin ruido			Conjunto de prueba con ruido		
	Precisión	Sensibilidad	Puntaje F1	Precisión	Sensibilidad	Puntaje F1
8	97.60%	99.19%	98.39%	96.03%	98.37%	97.19%
16	100.00%	99.16%	99.58%	96.69%	98.32%	97.50%
24	98.39%	98.39%	98.39%	97.50%	94.35%	95.90%
32	93.94%	92.54%	93.23%	94.17%	84.33%	88.98%
40	94.83%	95.65%	95.24%	96.36%	92.17%	94.22%
48	98.21%	98.21%	98.21%	96.52%	99.11%	97.80%
56	96.85%	100.00%	98.40%	95.28%	98.37%	96.80%
64	99.12%	99.12%	99.12%	94.02%	96.49%	95.24%
72	**91.67**%	95.28%	93.44%	**82.00**%	96.85%	88.81%
80	100.00%	97.46%	98.71%	99.12%	94.92%	96.97%
desconocido	93.18%	**89.13**%	**91.11**%	85.82%	**83.33**%	**84.56**%
silencio	100.00%	100.00%	100.00%	100.00%	95.12%	97.50%

Tabla 4.8: *Precisión, sensibilidad y puntaje F1 para el modelo de la tabla del 8 con arquitectura CRNN entrenado con el conjunto de entrenamiento aumentado*

4.2.8 Modelo de la tabla del 9



(a) *Conjunto de prueba sin ruido* **(b)** *Conjunto de prueba con ruido*

Figura 4.9: *Matrices de confusión para el modelo de la tabla del 9 con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado*

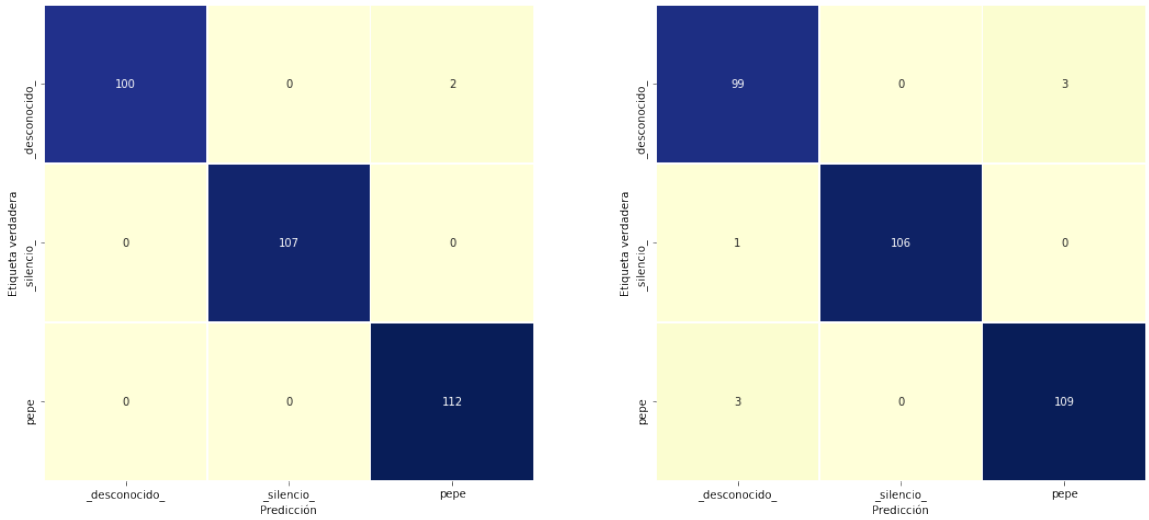
En la Fig. 4.9, se pueden observar las matrices de confusión para el modelo de la tabla del 9 con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado cuando se prueba con el conjunto sin ruido (Fig. 4.9a) y con ruido (Fig. 4.9b). Se puede apreciar que para ambos conjuntos de prueba, el modelo tiende a confundir en mayor medida el número 18 con la clase _desconocido_; y la clase _desconocido_ con el números 36 y 9. Para el conjunto de prueba con ruido, tambien se puede apreciar un aumento de los errores de clasificación al confundirse la clase _desconocido_ con el número 81.

En la Tabla 4.9 se muestra que para ambos conjuntos de prueba, los valores más bajos en las métricas corresponden a la clase _desconocido_, lo que indica que el modelo de la *tabla del 9* presenta dificultades al clasificar números que no se encuentren dentro de los resultados de la tabla de multiplicación del 9.

Etiqueta	**Conjunto de prueba sin ruido**			**Conjunto de prueba con ruido**		
	Precisión	**Sensibilidad**	**Puntaje F1**	**Precisión**	**Sensibilidad**	**Puntaje F1**
9	93.16%	96.46%	94.78%	92.24%	94.69%	93.45%
18	96.37%	95.28%	96.80%	96.75%	93.70%	95.20%
27	96.00%	99.17%	97.56%	92.06%	95.87%	93.93%
36	92.06%	95.87%	93.93%	86.36%	94.21%	90.12%
45	99.17%	100.00%	99.59%	98.28%	95.00%	96.61%
54	98.36%	100.00%	99.17%	96.75%	99.17%	97.94%
63	99.13%	99.13%	99. 13%	99.10%	95.65%	97.35%
72	96.95%	100.00%	98.45%	93.80%	95.28%	94.53%
81	97.56%	97.56%	97.56%	94.49%	97.56%	96.00%
90	98.29%	95.83%	97.05%	96.52%	92.50%	94.47%
desconocido	**85.04**%	**77.14**%	**80.90**%	**74.81**%	**70.00**%	**72.32**%
silencio	100.00%	100.00%	100.00%	99.17%	97.56%	98.36%

Tabla 4.9: *Precisión, sensibilidad y puntaje F1 para el modelo de la tabla del 9 con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado*

4.2.9 Modelo de activación



(a) *Conjunto de prueba sin ruido* **(b)** *Conjunto de prueba con ruido*

Figura 4.10: *Matrices de confusión para el modelo de activación con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado*

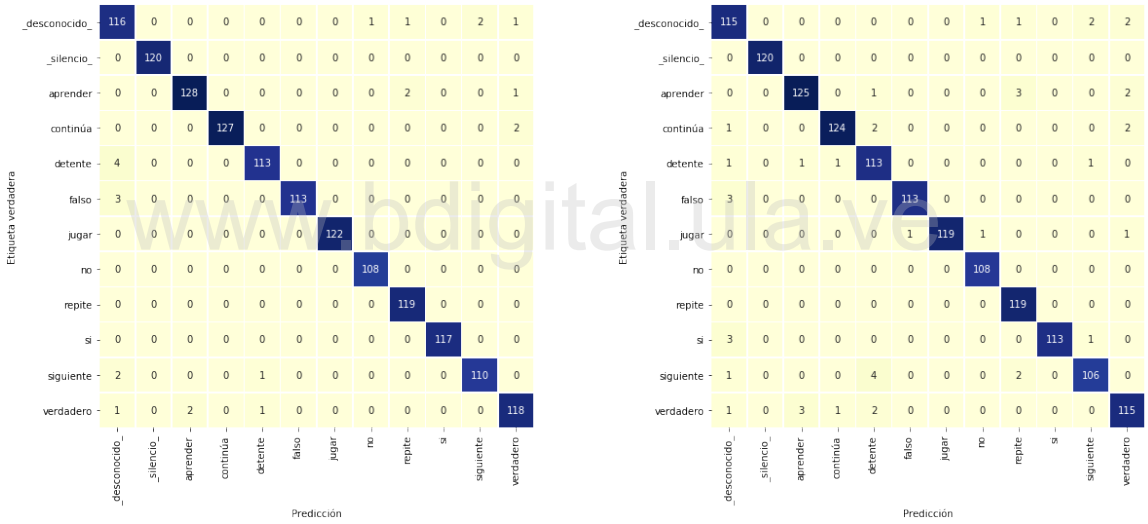
En la Fig. 4.10, se pueden observar las matrices de confusión para el modelo de *activación* con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado cuando se prueba con el conjunto sin ruido (Fig. 4.10a) y con ruido (Fig. 4.10b). Se puede apreciar que para el conjunto de prueba sin ruido el modelo confunde la clase _desconocido_ con la clase pepe. Para el conjunto de prueba con ruido este error de clasificación se mantiene. Además, se aprecia que el modelo confunde la clase pepe con la clase _desconocido_. Aunque el modelo clasifica erróneamente algunas muestras, para ambos conjuntos de prueba, el modelo logra un buen desempeño al clasificar correctamente cada una de la clases.

En la Tabla 4.10 se muestra que para ambos conjuntos de prueba, el valor más bajo de precisión corresponde a la clase pepe, mientras que los valores más bajos de sensibilidad y puntaje F1 corresponden a la clase _desconocido_.

Etiqueta	Conjunto de prueba sin ruido			Conjunto de prueba con ruido		
	Precisión	**Sensibilidad**	**Puntaje F1**	**Precisión**	**Sensibilidad**	**Puntaje F1**
pepe	**98.25**%	100.00%	99.12%	97.32%	97.32%	97.32%
desconocido	100.00%	**98.04**%	**99.01**%	**96.12**%	**97.0**6%	**96.59**%
silencio	100.00%	100.00%	100.00%	100.00%	99.07%	99.53%

Tabla 4.10: *Precisión, sensibilidad y puntaje F1 para el modelo de activación con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado*

4.2.10 Modelo de interacción



(a) *Conjunto de prueba sin ruido* **(b)** *Conjunto de prueba con ruido*

Figura 4.11: *Matrices de confusión para el modelo de interacción con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado*

En la Fig. 4.11, se pueden observar las matrices de confusión para el modelo de *interacción* con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado cuando se prueba con el conjunto sin ruido (Fig. 4.11a) y con ruido (Fig. 4.11b). Se puede apreciar que en general, para ambos conjuntos de prueba, el modelo logra un buen desempeño al clasificar correctamente cada una de la clases.

En la Tabla 4.11 se muestra que para el conjunto de prueba sin ruido, los valores más bajos en las métricas corresponden a la clase _desconocido_, mientras que para el conjunto de prueba con ruido, la clase siguiente presenta el valor más bajo de sensibilidad, y la clase _desconocido_ los valores más bajos de precisión y puntaje F1. Esto indica, que el modelo de interacción en condiciones ruidosas, tiende a clasificar otras clases como _desconocido_ y rechazar la clase siguiente clasificándola como otra clase.

Etiqueta	**Conjunto de prueba sin ruido**			**Conjunto de prueba con ruido**		
	Precisión	**Sensibilidad**	**Puntaje F1**	**Precisión**	**Sensibilidad**	**Puntaje F1**
siguiente	98.21%	97.35%	97.78%	96.36%	**93.81**%	95.07%
verdadero	96.72%	96.72%	96.72%	94.26%	94.26%	94.26%
aprender	98.46%	97.71%	98.08%	96.90%	95.42%	96.15%
continúa	100.00%	98.45%	99.22%	98.41%	96.12%	97.25%
detente	98.26%	96.58%	97.41%	92.62%	96.58%	94.56%
falso	100.00%	97.41%	98.69%	99.12%	97.41%	98.26%
jugar	100.00%	100.00%	100.00%	100.00%	97.54%	98.76%
no	99.08%	100.00%	99.54%	98.18%	100.00%	99.08%
repite	97.54%	100.00%	98.76%	95.20%	100.00%	97.54%
sí	100.00%	100.00%	100.00%	100.00%	96.58%	98.26%
desconocido	**92.06**%	**95.87**%	**93.93**%	**92.00**%	95.04%	**93.50**%
silencio	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Tabla 4.11: *Precisión, sensibilidad y puntaje F1 para el modelo de interacción con arquitectura GRU entrenado con el conjunto de entrenamiento aumentado*

4.3 Análisis de los resultados

Basándose en los resultados obtenidos en cada una de las pruebas realizadas sobre cada uno de los modelos implementados (ver Secciones 4.1 y 4.2), se pueden realizar las siguientes observaciones:

- Los resultados de las diferentes métricas obtenidas para los modelos entrenados con el conjunto de entrenamiento aumentado, muestran una evidente mejora con respecto al coste de los falsos positivos (precisión), falsos negativos (sensibilidad) y puntaje F1, cuando se realizan las pruebas con los conjuntos de pruebas sin ruido y con ruido. Lo que apunta a que el aumento de datos aplicado sí mejora el rendimiento de los modelos al mejorar las tasas de reconocimiento y la robustez al ruido de los modelos.

- Con el cálculo de métricas de exactitud y el micro promedio del área bajo la curva (AUC) de la curva de característica operativa del receptor (ROC), se observó que utilizando los modelos con arquitecturas GRU y CRNN entrenados con el conjunto de entrenamiento aumentado se obtuvieron los mejores resultados. Lo cual concuerda con los resultados que se muestran en la Tabla 3.13.

- De las 3 arquitecturas probadas, la que obtiene el menor rendimiento para cada uno de los modelos es la arquitectura CNN. Aunque este tipo de arquitectura logra explotar la correlación temporal y espectral local en las características del habla, las arquitecturas RNN y CRNN demuestran un mejor rendimiento ya que no solo explotan la relación temporal local de la señal de entrada, sino que también capturan la relación a largo plazo al usar celdas recurrentes.

- Los errores de clasificación presentes en los modelos de las tablas de multiplicar, se deben principalmente a que algunos dígitos son acústicamente muy similares entre sí. Un ejemplo de esto, es el caso del número 18 y el número 16 (ver sección 4.2.1), el número 45 y el número 35 (ver sección 4.2.4) o el número 32 con el número 72 (ver sección 4.2.7). Otra razón de los errores de clasificación se puede explicar teniendo en cuenta que, algunos números son de corta duración acústica, normalmente de unos pocos mili segundos de habla, como por ejemplo

el número 4 y el número 8 en la tabla del 2 (ver sección 4.2.1). Aunado a esto, los niños tienden a reemplazar un fonema por otro y pronunciar incorrectamente las números cuando hablan, aumentando aún más los errores de clasificación.

- La clase _desconocido_, es la que obtuvo los resultados más bajos para las métricas de precisión, sensibilidad y puntaje F1 para casi todos los modelos. Esto puede deberse al hecho de que esta clase está conformada por varias muestras distintas entre sí y que además pueden ser acústicamente similares a las otras clases. Esto dificulta a los modelos obtener una mejor representación para esta clase y aumenta los errores de clasificación.

- En general, los mejores resultados fueron obtenidos por los modelos de *activación* e *interacción*. Esto puede deberse al hecho de que el modelo de *activación* sólo clasifica entre un número reducido de clases, las cuales son muy distintas acústicamente. De igual forma, el modelo de *interacción*, al contrario que los modelos de las tablas que clasifican números, clasifica palabras clave muy distintas acústicamente. Esto permite a los modelos diferenciar entre clases más fácilmente y obtener un mejor rendimiento.

- Para las pruebas realizadas con el conjunto de prueba limpio, todos los modelos obtuvieron métricas perfectas para la clase _silencio_. Esto se puede explicar teniendo en cuenta que las muestras etiquetadas como _silencio_ en este conjunto de prueba son solo muestras totalmente silenciosas (valores 0), lo cual puede permitir al modelo clasificarlas de manera fácil.

Capítulo 5

Conclusiones y recomendaciones

Este proyecto de grado tuvo como objetivo general diseñar una interacción humano-robot para el aprendizaje de las tablas de multiplicación e implementar los modelos de reconocimiento del habla necesarios para llevar a cabo la interacción. Esta interacción fue diseñada con el objetivo de brindar una alternativa a las estrategias actuales para el aprendizaje de las tablas de multiplicación, que permita cubrir las desventajas y aprovechar las ventajas de éstas, y que además, tome en consideración los elementos más importantes presentes al usar robots sociales en el contexto de la educación matemática.

En aras de que la robótica social continúe progresando hacia entornos del mundo real, en escenarios donde se interactúe con niños, se tomó en consideración la comunicación verbal con el robot. Por lo cual, se diseñaron e implementaron 10 modelos de reconocimiento de palabras clave, para llevar a cabo la interacción diseñada de manera verbal. Para esto, se construyó un corpus de audio infantil denominado LaSDAICVI, para el entrenamiento y evaluación de cada uno de los modelos de reconocimiento de palabras clave implementados. Adicionalmente, fueron realizadas diferentes implementaciones y pruebas, que sirvieron para comparar y evaluar el rendimiento de los distintos modelos. A continuación, se presentan las conclusiones, aportes, recomendaciones y trabajos futuros de este proyecto de grado.

5.1 Conclusiones

En este proyecto de grado, se diseñó una interacción humano-robot para el aprendizaje de las tablas de multiplicación. El diseño de esta interacción nace de la de necesidad de brindar una alternativa que puede cubrir las desventajas más importantes presentes en las estrategias actuales para el aprendizaje de las tablas de multiplicación. Para el diseño de esta interacción, se consultaron varias estrategias que se aplican actualmente, con la finalidad de determinar cuáles eran las ventajas y desventajas que éstas poseían y que pudieran ser considerados en la interacción humano-robot a diseñar. De igual forma, se realizó una revisión de trabajos relacionados con el uso de robots sociales en el contexto de la educación matemática, con el objetivo de encontrar los elementos que éstos aplican, y que permiten facilitar y aumentar la motivación de los niños durante el aprendizaje de un tema matemático, para posteriormente integrarlos a la interacción humano-robot para el aprendizaje de las tablas de multiplicación diseñada. La interacción humano-robot diseñada, constó de 3 etapas donde un robot lleva a cabo dos juegos de preguntas y respuestas sobre las tablas de multiplicación junto con un niño. En la interacción, el robot deberá adaptarse a las debilidades del niño con las tablas de multiplicación para practicar las tablas que más se le dificultan.

Una parte importante de la estrategia de interacción humano-robot que fue diseñada, es que se agregó la capacidad de que los niños puedan interactuar a través de la voz con el robot. El reconocimiento del habla desarrolla un papel importante en la robótica social, ya que permite ofrecer una forma de comunicación con los robots mucho más natural e intuitiva, similar a la existente entre los humanos. Sin embargo, en la actualidad la mayoría de los corpus de audio destinados al entrenamiento y evaluación de los modelos de reconocimiento del habla, se centra principalmente en el habla de personas adultas. Ésto plantea un desafío, debido a la carencia de corpus de audio infantil para el entrenamiento y evaluación de modelos de reconocimiento del habla infantil. Por tal motivo, en este proyecto de grado se construyó el corpus de audio infantil en español "LaSDAI Comandos de Voz Infantil" (LaSDAICVI) con la intención de que pueda ser usado para el entrenamiento y evaluación de modelos de reconocimiento de palabras clave en español a través del habla infantil. Para esto, se realizó una investigación de corpus de audio infantiles, con el objetivo de estudiar

los factores a considerar al momento de diseñar y construir nuestro corpus de audio con niños. Parte de su diseño consistió en la definición de un conjunto de palabras y números, las cuales consistieron en la serie de números del 0 al 9, junto con los números resultantes en las operaciones de las tablas de multiplicación del 2 al 9, además de 18 palabras necesarias que servirán como comandos de voz para desarrollar la interacción diseñada. LaSDAICVI consta de un total de 29061 muestras de audio, las cuales fueron compiladas de un total de 41 niños matriculados en escuelas primarias, pertenecientes a los grados tercero a sexto, con edades comprendidas entre los 8 y 11 años.

Con la intención de abordar el problema del reconocimiento del habla de una forma eficiente, decidimos dividirlo en problemas de menor complejidad. Por lo tanto, optamos por un enfoque de reconocimiento de palabras clave. Haciendo uso del corpus de audio infantil LaSDAICVI, se diseñaron e implementaron 10 modelos de reconocimiento de palabras clave; 1 modelo para cada tabla de multiplicación del 2 al 9, que reconoce los números presentes en los resultados de las tablas de multiplicación (para un total de 8 modelos); 1 modelo que reconoce los comandos y palabras requeridas durante la interacción; y 1 modelo que reconoce la palabra de activación "Pepe". Fueron probadas 3 arquitecturas de redes neuronales diferentes para cada modelo: redes neuronales convolucionales, redes neuronales recurrentes y redes neuronales convolucionales recurrentes, seleccionadas a partir de investigaciones previas en el área del reconocimiento de palabras clave.

Para cada uno de los modelos entrenados se calcularon diferentes métricas para seleccionar aquellos con el mejor rendimiento y realizar un análisis más profundo de las particularidades de cada modelo. Los resultados obtenidos mostraron que los modelos con arquitecturas de redes neuronales recurrentes y convolucionales recurrentes obtuvieron los mejores resultados, ya que estos explotan tanto la relación temporal local como a largo plazo de las señales de audio al utilizar celdas recurrentes. Además, se evidenció que aquellos modelos entrenados con el conjunto de entrenamiento con aumento de datos mostraban una mejora sustancial con respecto a aquellos entrenados con el conjunto de entrenamiento sin aumento de datos, demostrando que el aumento de datos aplicado mejora las tasas de reconocimiento y la robustez al ruido de los modelos. Finalmente, observamos que los errores de clasificación presentes en los modelos de las

tablas de multiplicar, se debían principalmente similitudes o cortas duraciones acústicas de los números; y al hecho de que los niños tienden a reemplazar un fonema por otro y pronunciar incorrectamente las números cuando hablan, aumentando aún más los errores de clasificación.

En conclusión, los objetivos planteados en el capítulo 1 fueron alcanzados tras haber diseñado la interacción humano-robot para el aprendizaje de las tablas de multiplicación, construido el corpus de audio infantil para el entrenamiento y evaluación de los modelos de reconocimiento de habla infantil y haber diseñado e implementado los modelos para el reconocimiento del habla según la interacción humano-robot diseñada.

5.2 Aportes

Las principales contribuciones de este proyecto de grado son las siguientes:

- Se realizó una revisión sobre las estrategias utilizadas para el aprendizaje de la tablas de multiplicación.
- Se realizó una revisión sobre los robots sociales en el contexto de la educación matemática.
- Se realizó una revisión sobre los corpus de audios disponibles actualmente para el reconocimiento del habla infantil.
- Se realizó una revisión sobre los modelos para el reconocimiento del habla infantil.
- Se realizó la construcción de un corpus de audio infantil para el reconocimiento de palabras clave, que hasta momento de realización este proyecto de grado, sería el primer corpus de audio infantil en español para este tipo de aplicaciones.
- Se diseñó una interacción humano-robot para el aprendizaje de las tablas de multiplicación que permite adaptarse a las dificultades que los niños presentan con las tablas de multiplicación

- Se diseñaron e implementaron 10 modelos de reconocimientos de palabras clave, para permitir el reconocimiento del habla en la interacción humano-robot diseñada.
- Se evaluaron varios tipos de arquitecturas de redes neuronales en los modelos de reconocimiento de palabras clave para determinar cual generaba los mejores resultados.
- Se evaluó el aumento de datos sobre el corpus de audio LaSDAICVI y su efecto sobre los modelos de reconocimiento de palabras clave.

5.3 Recomendaciones

A continuación se presentan algunas recomendaciones sobre la interacción humano-robot diseñada y los modelos de reconocimiento de palabras clave desarrollados.

- A nivel lingüístico, los niños pueden reemplazar un fonema por otro y son más propensos a usar palabras imaginarias, frases gramaticalmente incorrectas y pronunciar incorrectamente las palabras mientras estén interactuando con un robot. Esto puede aumentar los errores de clasificación en los modelos de reconocimiento de palabras clave. Por tal motivo, es recomendable acompañar la interacción con otra técnica como la del "Mago de Oz" para cubrir los errores cometidos por el robot y brindar una experiencia de usuario más agradable.
- Por lo general, las condiciones en las cuales interactúa un robot un social son bastante controladas. En la interacción humano-robot diseñada fue tomado en consideración este hecho. Por lo tanto, se recomienda que una vez implementada la interacción en un robot, se prepare con anticipación al niño para posibles problemas en el comportamiento del robot, por ejemplo, si el robot se equivoca al marcar una respuesta como correcta o incorrecta. De esta forma, se evitará que el participante no esté seguro de lo que deba hacer en tal situación.

- Una de las principales desventajas en el reconocimiento de patrones mediante el audio, es que éste es muy susceptible a las condiciones de ambiente. Aunque se aplicó aumento de datos añadiendo ruido a las muestras de entrenamiento para permitir que los modelos implementados fueran más robustos al ruido, el ruido ambiental es impredecible. Por lo tanto, para obtener los mejores resultados, las condiciones ambientales deben ser similares a las utilizadas para el entrenamiento de los modelos de reconocimiento de palabras clave descritos en este proyecto de grado.

5.4 Trabajos Futuros

A continuación se presentan trabajos futuros que surgieron a partir de este proyecto de grado:

- Desarrollar un modelo adaptativo utilizando aprendizaje reforzado que permita aprender las deficiencias de los niños con las tablas de multiplicación para que pueda ser integrado a la interacción humano-robot diseñada.
- Desarrollar un modelo de reconocimiento facial utilizando aprendizaje profundo que permita reconocer a los niños y pueda ser integrado a la interacción humano-robot diseñada.
- Implementar la interacción humano-robot diseñada en un robot social para probar su efecto en el aprendizaje de las tablas de multiplicación de los niños.

Bibliografía

[1] T. L. Solomon and J. Mighton, "Developing mathematical fluency: A strategy to help children learn their multiplication facts," *Perspectives on Language and Literacy*, vol. 43, no. 1, pp. 31–34, 2017.

[2] B. Allen-Lyall, "Helping students to automatize multiplication facts: A pilot study," *International Electronic Journal of Elementary Education*, vol. 10, no. 4, pp. 391–396, 2018.

[3] M. M. R. Hernández, J. L. G. Fernández, and R. R. Bastante, "Las tablas de multiplicar con sabor a juego. Recursos didácticos," *Números: Revista de didáctica de las matemáticas*, no. 90, pp. 7–19, 2015.

[4] R. A. Reina and K. V. Ramírez, "¿Memorizar las tablas de multiplicar garantiza el aprendizaje y la comprensión en los niños?" *Revista Ejes*, vol. 1, no. 1, pp. 18–21, 2013.

[5] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," *Science robotics*, vol. 3, no. 21, p. eaat5954, 2018.

[6] N. F. Chen, R. Tong, D. Wee, P. X. Lee, B. Ma, and H. Li, "Singakids-mandarin: Speech corpus of singaporean children speaking mandarin chinese." in *Interspeech*, 2016, pp. 1545–1549.

[7] H. Liao, G. Pundak, O. Siohan, M. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani, "Large vocabulary automatic speech recognition for children," in *INTERSPEECH*, 2015, pp. 1611–1615.

[8] A. Hämäläinen, S. Rodrigues, A. Júdice, S. M. Silva, A. Calado, F. M. Pinto, and M. S. Dias, "The cng corpus of european portuguese children's speech," in *International Conference on Text, Speech and Dialogue*. Springer, 2013, pp. 544–551.

[9] D. M. G. Rivera, "Software educativo como ejercitador y herramienta didáctica en comparación con el ábaco abierto para aumentar el aprendizaje de las tablas de multiplicar en niños y niñas de segundo de primaria," *Revista Electrónica de Educación y Psicología*, vol. 2, no. 4, 2006.

[10] O. A. Usuga Macias, "Diseño de una unidad didáctica para la enseñanza-aprendizaje de la multiplicación de números naturales en el grado tercero de la institución educativa Antonio Derka Santo Domingo del municipio de Medellín," Master's thesis, Universidad Nacional de Colombia-Sede Medellín, 2014.

[11] M. C. Barrera Barrera, "Estrategias metodológicas y su incidencia en el aprendizaje de las tablas de multiplicar en los niños/as de quinto grado paralelo a de educación básica del centro de educación general básica Manuela Espejo del Cantón Ambato provincia de Tungurahua," B.S. thesis, Universidad Técnica de Ambato, 2013.

[12] A. Borrero Monge, "Juegos y materiales manipulativos como recurso didáctico para enseñar las tablas de multiplicar," B.S. thesis, Universidad de Sevilla, 2018.

[13] C. L. Muñoz Ortíz *et al.*, "Estrategias didácticas para desarrollar el aprendizaje significativo de las tablas de multiplicar en niños del grado 3–b de la institución educativa Jose Holguín Garcés–sede Ana María de Lloreda," B.S. thesis, Universidad de La Sabana, 2012.

[14] N. Rodrigo-Huete, "Enseñar a multiplicar mediante el juego y el aprendizaje cooperativo," B.S. thesis, Universidad Internacional de La Rioja, 2017.

[15] J. I. Gomez Ruiz and L. C. Perez Ozuna, "Diseño e implementación de un software educativo que permite desarrollar habilidades y destrezas para el manejo

de las tablas de multiplicar en los estudiantes del grado 3º de la institución educativa Rafael Nuñez sede Bolivar del municipio de Sincelejo," Master's thesis, Universidad Francisco de Paula Santander Ocaña, 2014.

[16] W. J. Forero Martínez, E. A. Granados Velásquez, P. Sierra, and F. José, "Software educativo ludo-pedagógico para solucionar problemas en la enseñanza y aprendizaje de las tablas de multiplicar con estudiantes del grado tercero de la institución educativa técnica La Integrada sede Barrio Nuevo del municipio de San Pablo Sur de Bolívar," Master's thesis, Fundación Universitaria Los Libertadores, 2015.

[17] S. Smith, A. W. Steele, J. du Toit, and M. Conning, "Development of an educational tool to teach primary school pupils multiplication tables," in *2015 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*. IEEE, 2015, pp. 19–23.

[18] E. Vrochidou, A. Najoua, C. Lytridis, M. Salonidis, V. Ferelis, and G. A. Papakostas, "Social robot nao as a self-regulating didactic mediator: a case study of teaching/learning numeracy," in *2018 26th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*. IEEE, 2018, pp. 1–5.

[19] K. R. Liles and J. M. Beer, "Rural minority students' perceptions of Ms. an, the robot teaching assistant, as a social teaching tool," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 59, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2015, pp. 372–376.

[20] A. Ramachandran, A. Litoiu, and B. Scassellati, "Shaping productive help-seeking behavior during robot-child tutoring interactions," in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 2016, pp. 247–254.

[21] A. Ramachandran, C.-M. Huang, and B. Scassellati, "Give me a break!: Personalized timing strategies to promote learning in robot-child tutoring," in

Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. ACM, 2017, pp. 146–155.

[22] K. V. Hindriks and S. Liebens, "A robot math tutor that gives feedback," in *International Conference on Social Robotics*. Springer, 2019, pp. 601–610.

[23] E. A. Konijn and J. F. Hoorn, "Robot tutor and pupils' educational ability: Teaching the times tables," *Computers & Education*, vol. 157, p. 103970, 2020.

[24] F. Claus, H. Gamboa Rosales, R. Petrick, H.-U. Hain, and R. Hoffmann, "A survey about asr for children," in *Speech and Language Technology in Education*, 2013.

[25] M. Russell and S. D'Arcy, "Challenges for computer recognition of children's speech," in *Workshop on Speech and Language Technology in Education*, 2007.

[26] A. Potamianos and S. Narayanan, "A review of the acoustic and linguistic properties of children's speech," in *2007 IEEE 9th Workshop on Multimedia Signal Processing*. IEEE, 2007, pp. 22–25.

[27] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris *et al.*, "Automatic speech recognition and speech variability: A review," *Speech communication*, vol. 49, no. 10-11, pp. 763–786, 2007.

[28] E. Booth, J. Carns, C. Kennington, and N. Rafla, "Evaluating and improving child-directed automatic speech recognition," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 6340–6345.

[29] S. S. Gray, D. Willett, J. Lu, J. Pinto, P. Maergner, and N. Bodenstab, "Child automatic speech recognition for us english: child interaction with living-room-electronic-devices." in *WOCCI*, 2014, pp. 21–26.

[30] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, no. 10-11, pp. 847–860, 2007.

[31] G. Yeung and A. Alwan, "On the difficulties of automatic speech recognition for kindergarten-aged children." in *INTERSPEECH*, 2018, pp. 1661–1665.

[32] A. Kazemzadeh, H. You, M. Iseli, B. Jones, X. Cui, M. Heritage, P. Price, E. Anderson, S. Narayanan, and A. Alwan, "Tball data collection: the making of a young children's speech corpus," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[33] P. B. Ramteke, S. Supanekar, P. Hegde, H. Nelson, V. Aithal, and S. Koolagudi, "Nitk kids' speech corpus," *emotion*, vol. 491, pp. 4–15, 2019.

[34] L. Cleuren, J. Duchateau, P. Ghesquiere *et al.*, "Children's oral reading corpus (chorec): description and assessment of annotator agreement," *LREC 2008 Proceedings*, pp. 998–1005, 2008.

[35] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[36] K. Shobaki, J.-P. Hosom, and R. A. Cole, "The ogi kids' speech corpus and recognizers," in *Sixth International Conference on Spoken Language Processing*, 2000.

[37] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The htk book," *Cambridge university engineering department*, vol. 3, no. 175, p. 12, 2002.

[38] S. Sutton, R. A. Cole, J. d. Villiers, J. Schalkwyk, P. Vermeulen, M. W. Macon, Y. Yan, E. Kaiser, B. Rundle, K. Shobaki *et al.*, "Universal speech tools: The cslu toolkit," in *Fifth International Conference on Spoken Language Processing*, 1998.

[39] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[40] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *STIN*, vol. 93, p. 27403, 1993.

[41] F. D. Rahman, N. Mohamed, M. B. Mustafa, and S. S. Salim, "Automatic speech recognition system for malay speaking children," in *2014 Third ICT International Student Project Conference (ICT-ISPC)*. IEEE, 2014, pp. 79–82.

[42] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, "Improving children's speech recognition through out-of-domain data augmentation." in *Interspeech*, 2016, pp. 1598–1602.

[43] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, "The pf_star children's speech corpus," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[44] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "Wsjcamo: a british english speech corpus for large vocabulary continuous speech recognition," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1995, pp. 81–84.

[45] S. M. D'Arcy, M. J. Russell, S. R. Browning, and M. J. Tomlinson, "The accents of the british isles (abi) corpus," *Proceedings Modélisations pour l'Identification des Langues*, pp. 115–119, 2004.

[46] M. Qian, I. McLoughlin, W. Quo, and L. Dai, "Mismatched training data enhancement for automatic recognition of children's speech using dnn-hmm," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2016, pp. 1–5.

[47] S. Watson and A. Coy, "Jamlit: A corpus of jamaican standard english for automatic speech recognition of children's speech." in *SLTU*, 2018, pp. 243–247.

[48] S. Fernando, R. K. Moore, D. Cameron, E. C. Collins, A. Millings, A. J. Sharkey, and T. J. Prescott, "Automatic recognition of child speech for robotic applications in noisy environments," *arXiv preprint arXiv:1611.02695*, 2016.

[49] H. Christensen, J. Barker, N. Ma, and P. D. Green, "The chime corpus: a resource and a challenge for computational hearing in multisource environments," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[50] M. Wöllmer, B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Tandem decoding of children's speech for keyword detection in a child-robot interaction scenario," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 7, no. 4, pp. 1–22, 2011.

[51] A. Batliner, S. Steidl, and E. Nöth, "Releasing a thoroughly annotated and processed spontaneous emotional database: the fau aibo emotion corpus," in *Proc. of a Satellite Workshop of LREC*, vol. 28, 2008.

[52] H. Sundar, J. F. Lehman, and R. Singh, "Keyword spotting in multi-player voice driven games for children," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[53] J. F. Lehman, N. Wolfe, and A. Pereira, "Multi-party language interaction in a fast-paced game using multi-keyword spotting," in *International Conference on Intelligent Virtual Agents*. Springer, 2016, pp. 331–340.

[54] J. F. Lehman and S. Al Moubayed, "Mole madness–a multi-child, fast-paced, speech-controlled game," in *2015 AAAI Spring Symposium Series*, 2015.

[55] R. Harvey-Swanston, "I was good at my times tables when i was nine, now i can't remember them": Learning multiplication facts with conceptual understanding," *Mathematics Teaching, December 2017*, vol. 259, pp. 20–22, 2017.

[56] L. Lotero Botero, E. Andrade Londoño, and L. Andrade Lotero, "La crisis de la multiplicación: Una propuesta para la estructuración conceptual," *Voces y*

silencios. Revista latinoamericana de educación, vol. 2, no. especial, pp. 38–64, 2011.

[57] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme, "Child speech recognition in human-robot interaction: evaluations and recommendations," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 82–90.

[58] Y. Yu, "Research on speech recognition technology and its application," in *2012 International Conference on Computer Science and Electronics Engineering*, vol. 1. IEEE, 2012, pp. 306–309.

[59] G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal, "Affective personalization of a social robot tutor for children's second language skills," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[60] D. Hood, S. Lemaignan, and P. Dillenbourg, "When children teach a robot to write: An autonomous teachable humanoid which uses simulated handwriting," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2015, pp. 83–90.

[61] M. Fridin, "Storytelling by a kindergarten social assistive robot: A tool for constructive learning in preschool education," *Computers & education*, vol. 70, pp. 53–64, 2014.

[62] G. Shuzhi and M. Maja, "Preface," in *International Journal of Social Robotics*, vol. 1, no. 1. Springer, 2009, pp. 1–2.

[63] M. M. de Graaf, S. B. Allouch, and J. A. van Dijk, "Long-term evaluation of a social robot in real homes," *Interaction studies*, vol. 17, no. 3, pp. 462–491, 2016.

[64] S. Rossi, M. Larafa, and M. Ruocco, "Emotional and behavioural distraction by a social robot for children anxiety reduction during vaccination," *International Journal of Social Robotics*, vol. 12, no. 3, pp. 765–777, 2020.

[65] C. Bartneck and J. Forlizzi, "A design-centred framework for social human-robot interaction," in *RO-MAN 2004. 13th IEEE international workshop on robot and human interactive communication (IEEE Catalog No. 04TH8759)*. IEEE, 2004, pp. 591–594.

[66] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots: Concepts, design and applications," 2002.

[67] F. Hegel, S. Krach, T. Kircher, B. Wrede, and G. Sagerer, "Understanding social robots: A user study on anthropomorphism," in *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2008, pp. 574–579.

[68] C. L. Breazeal, *Designing sociable robots*. MIT press, 2002.

[69] T. Fong, C. Thorpe, and C. Baur, "Collaboration, dialogue, human-robot interaction," in *Robotics Research*. Springer, 2003, pp. 255–266.

[70] J. Pérez, J. Aguilar, and E. Dapena, "Mihr: A human-robot interaction model," *IEEE Latin America Transactions*, vol. 18, no. 9, pp. 1521–1529, 2020.

[71] J. Pérez, J. Aguilar, and E. Dapena, "Mihh: Un modelo de interacción humano-humano," *Revista Venezolana de Computación*, vol. 5, no. 1, pp. 10–19, 2018.

[72] N. Morán, "Módulo reconfigurable de reconocimiento para la interacción humano-robot," B.S. thesis, Universidad de Los Andes, 2019.

[73] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.

[74] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[75] A. Saran, S. Majumdar, E. S. Short, A. Thomaz, and S. Niekum, "Human gaze following for human-robot interaction," in *2018 IEEE/RSJ International*

Conference on Intelligent Robots and Systems (IROS). IEEE, 2018, pp. 8615–8621.

[76] M. Atzeni and D. R. Recupero, "Deep learning and sentiment analysis for human-robot interaction," in *European Semantic Web Conference*. Springer, 2018, pp. 14–18.

[77] J. Deng, G. Pang, Z. Zhang, Z. Pang, H. Yang, and G. Yang, "cgan based facial expression recognition for human-robot interaction," *IEEE Access*, vol. 7, pp. 9848–9859, 2019.

[78] M. Suguitan, M. Bretan, and G. Hoffman, "Affective robot movement generation using cyclegans," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 534–535.

[79] Y. LeCun *et al.*, "Generalization and network design strategies," *Connectionism in perspective*, vol. 19, pp. 143–155, 1989.

[80] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[81] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[82] A. Mechelli and S. Vieira, *Machine learning: methods and applications to brain disorders*. Academic Press, 2019.

[83] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[84] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention lstm networks," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1586–1599, 2017.

[85] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[86] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee, 2013, pp. 6645–6649.

[87] S. El Hihi and Y. Bengio, "Hierarchical recurrent neural networks for long-term dependencies." in *Nips*, vol. 409. Citeseer, 1995.

[88] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[89] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[90] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 1422–1432.

[91] Z. Zuo, B. Shuai, G. Wang, X. Liu, X. Wang, B. Wang, and Y. Chen, "Convolutional recurrent neural networks: Learning spatial dependencies for image representation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 18–26.

[92] S. O. Arik, M. Kliegl, R. Child, J. Hestness, A. Gibiansky, C. Fougner, R. Prenger, and A. Coates, "Convolutional recurrent neural networks for small-footprint keyword spotting," *arXiv preprint arXiv:1703.05390*, 2017.

[93] Y. Zhang, N. Suda, L. Lai, and V. Chandra, "Hello edge: Keyword spotting on microcontrollers," *arXiv preprint arXiv:1711.07128*, 2017.

[94] O. Rybakov, N. Kononenko, N. Subrahmanya, M. Visontai, and S. Laurenzo, "Streaming keyword spotting on mobile devices," *arXiv preprint arXiv:2005.06720*, 2020.

[95] T. Lan, S. Aryal, B. Ahmed, K. Ballard, and R. Gutierrez-Osuna, "Flappy voice: an interactive game for childhood apraxia of speech therapy," in *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play*, 2014, pp. 429–430.

[96] H. Jung, H. J. Kim, S. So, J. Kim, and C. Oh, "Turtletalk: an educational programming game for children with voice user interface," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–6.

[97] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 65–78, 2002.

[98] S. Fernández, A. Graves, and J. Schmidhuber, "An application of recurrent neural networks to discriminative keyword spotting," in *International Conference on Artificial Neural Networks*. Springer, 2007, pp. 220–229.

[99] R. Tang and J. Lin, "Deep residual learning for small-footprint keyword spotting," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5484–5488.

[100] G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4087–4091.

[101] T. N. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[102] S. Choi, S. Seo, B. Shin, H. Byun, M. Kersner, B. Kim, D. Kim, and S. Ha, "Temporal convolution for real-time keyword spotting on mobile devices," *arXiv preprint arXiv:1904.03814*, 2019.

[103] I. Arroyo, B. P. Woolf, W. Burelson, K. Muldner, D. Rai, and M. Tai, "A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect," *International Journal of Artificial Intelligence in Education*, vol. 24, no. 4, pp. 387–426, 2014.

[104] R. Ros, M. Nalin, R. Wood, P. Baxter, R. Looije, Y. Demiris, T. Belpaeme, A. Giusti, and C. Pozzi, "Child-robot interaction in the wild: advice to the aspiring experimenter," in *Proceedings of the 13th international conference on multimodal interfaces*, 2011, pp. 335–342.

[105] E. L. Deci and R. M. Ryan, "Intrinsic motivation," *The corsini encyclopedia of psychology*, pp. 1–2, 2010.

[106] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[107] T. Belpaeme, P. Baxter, J. De Greeff, J. Kennedy, R. Read, R. Looije, M. Neerincx, I. Baroni, and M. C. Zelati, "Child-robot interaction: Perspectives and challenges," in *International Conference on Social Robotics*. Springer, 2013, pp. 452–459.

[108] M. Zeng and N. Xiao, "Effective combination of densenet and bilstm for keyword spotting," *IEEE Access*, vol. 7, pp. 10 767–10 775, 2019.

[109] J. M. Ramirez, A. Montalvo, and J. R. Calvo, "A survey of the effects of data augmentation for automatic speech recognition systems," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2019, pp. 669–678.

[110] S. Shahnawazuddin, K. Maity, and G. Pradhan, "Improving the performance of keyword spotting system for children's speech through prosody modification," *Digital Signal Processing*, vol. 86, pp. 11–18, 2019.

[111] T. Giannakopoulos and A. Pikrakis, *Introduction to audio analysis: a MATLAB® approach*. Academic Press, 2014.

[112] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.

[113] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.