UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

GRADO EN INGENIERÍA BIOMÉDICA

TRABAJO FIN DE GRADO

DISEÑO E IMPLEMENTACIÓN DE UN SISTEMA DE GENERACIÓN DE VÍDEOS PARA LA REHABILITACIÓN DE PACIENTES CON AFASIA MEDIANTE EL USO DE INTELIGENCIA ARTIFICIAL GENERATIVA

NATALIA ESTEBAN DE NICOLÁS

2025

GRADO EN INGENIERÍA BIOMÉDICA

TRABAJO FIN DE GRADO

Título: Diseño e implementación de un sistema de generación de vídeos para la rehabilitación de pacientes con afasia mediante el uso de inteligencia artificial generativa.

Autor: Dña. Natalia Esteban de Nicolás

Tutor: Dña. Blanca Fuentes Gimeno

Ponente: D. Álvaro Gutiérrez Martín

Departamento: Laboratorio de Robótica y Control, ETSIT, UPM

MIEMBROS DEL TRIBUNAL

Presidente: D. ……

Vocal: D. ……..

Secretario: D. ……..

Suplente: D. ……..

Los miembros del tribunal arriba nombrados acuerdan otorgar la calificación de: ……

Madrid, a de de 20…

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

GRADO EN INGENIERÍA BIOMÉDICA

TRABAJO FIN DE GRADO

DISEÑO E IMPLEMENTACIÓN DE UN SISTEMA DE GENERACIÓN DE VÍDEOS PARA LA REHABILITACIÓN DE PACIENTES CON AFASIA MEDIANTE EL USO DE INTELIGENCIA ARTIFICIAL GENERATIVA

NATALIA ESTEBAN DE NICOLÁS

2025

RESUMEN

La afasia es un trastorno del lenguaje que compromete tanto la expresión como la comprensión verbal de las personas que la padecen. Entre sus causas principales se encuentra el ictus, una de las enfermedades neurológicas más graves, afectando anualmente a miles de personas en nuestro país y primera causa de discapacidad adquirida en adultos. La afasia interfiere considerablemente en la calidad de vida de los pacientes, quienes enfrentan barreras comunicativas que conllevan al aislamiento social y repercuten en su salud emocional. Ante esta situación, y con el objetivo de mejorar y acelerar el pronóstico de esta condición, surge la necesidad de implementar nuevas técnicas de rehabilitación del lenguaje complementarias a las actuales.

Este Trabajo de Fin de Grado presenta el diseño e implementación de un sistema de creación de vídeos personalizados para la rehabilitación de pacientes con afasia mediante el uso de inteligencia artificial (IA) generativa. Inspirado en el proyecto DULCINEA, que utiliza el doblaje de escenas pregrabadas de series de televisión para mejorar las habilidades comunicativas, este sistema busca superar la dependencia de contenido externo mediante la generación automática de vídeos contextualizados y personalizados a las necesidades de los pacientes.

El sistema integra tres modelos de IA generativa: Mochi 1 para la creación de personajes y escenarios, Azure Speech para la generación de voces, y Wav2Lip para la sincronización labial. Esta combinación permite generar escenas que simulan situaciones cotidianas y ofrecen un enfoque inmersivo y motivador para el paciente. El alto grado de personalización que ofrece, permite a los pacientes y terapeutas definir tanto los escenarios como las frases que desean practicar. Sin embargo, durante la elaboración de este proyecto se han identificado grandes limitaciones técnicas, como los altos requerimientos computacionales. Se ha cargado el modelo de baja resolución de Mochi 1 en un entorno de Google Colab Pro con acceso a hardware avanzado. No obstante, la calidad de los contenidos generados no es suficiente e incluso, en algunos casos, no se pueden identificar bien las caras de los personajes, imposibilitando la sincronización labial con Wav2Lip. De esta manera, se ha optado por generar demos con vídeos descargados directamente del playground online de Mochi 1, los cuales fueron posteriormente importados al sistema de generación automático de escenas, donde se generaron las voces y se realizó la sincronización labial. Este enfoque se realizó con el objetivo de visualizar cómo podrían ser los resultados finales si se contase con los recursos técnicos necesarios para generar vídeos de alta calidad de manera directa. Aun así, se han encontrado ligeras desincronizaciones entre el audio y los movimientos labiales.

En conclusión, este proyecto representa un avance significativo hacia la implementación de la IA generativa en el ámbito clínico. Aunque se encuentra en una etapa inicial, sienta las bases para futuros proyectos que podrían hacer frente a las limitaciones actuales y desarrollar un sistema de generación de vídeos más robusto y eficiente.

PALABRAS CLAVE

Afasia, rehabilitación, ictus, inteligencia artificial generativa, sincronización labial, vídeos personalizados, doblaje, Mochi 1, Azure Speech, Wav2Lip, Google Colab Pro.

SUMMARY

Aphasia is a language disorder that compromises both the verbal expression and comprehension of people who suffer from it. Among its main causes is stroke, one of the most serious neurological diseases, affecting thousands of people in our country every year and the leading cause of acquired disability in adults. Aphasia interferes considerably in the quality of life of patients, who face communication barriers that lead to social isolation and have repercussions on their emotional health. Given this situation, and with the aim of improving and accelerating the prognosis of this condition, there is a need to implement new language rehabilitation techniques to complement current ones.

This bachelor's Thesis introduces the design and implementation of a system for the creation of personalised videos for the rehabilitation of patients with aphasia using generative artificial intelligence (AI). Inspired by the DULCINEA project, which uses the dubbing of pre-recorded scenes from television series to improve communication skills, this system aims to overcome the dependency on external content by automatically generating contextualised videos personalised to the needs of patients.

The system integrates three generative AI models: Mochi 1 for character and scenario creation, Azure Speech for voice generation, and Wav2Lip for lip-synchronisation. This combination allows generated scenes to simulate real-life situations and offer an immersive and motivating approach for the patient. The high degree of customisation offered allows patients and therapists to define both the scenarios and the sentences they would like to practice. However, during the development of this project, major technical limitations have been identified, such as high computational requirements. The low-resolution Mochi 1 model has been loaded in a Google Colab Pro environment with access to advanced hardware. However, the quality of the generated content is not sufficient, and, in some cases, the faces of the characters cannot be identified properly, making lip-synchronisation with Wav2Lip impossible. Therefore, we opted to generate demos with videos downloaded directly from the Mochi 1 online playground, which were subsequently imported into the automatic scene generation system, where the voices were generated and the lip-sync was performed. This approach was taken in order to visualise what the final results could be like if the necessary technical resources were available to generate high quality videos directly. Even so, slight out-of-sync have been found between audio and lip movements.

In conclusion, this project represents a significant advance towards the implementation of generative AI in the clinical domain. Although it is still at an early stage, it lays the groundwork for future projects that could address current limitations and develop a more robust and efficient video generation system.

KEYWORDS

Aphasia, rehabilitation, stroke, generative artificial intelligence, Lip Sync, personalized videos, dubbing, Mochi 1, Azure Speech, Wav2Lip, Google Colab Pro.

ÍNDICE DEL CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN Y OBJETIVOS

1.1. INTRODUCCIÓN

El lenguaje es un sistema convencional de comunicación complejo y dinámico que proporciona al ser humano la capacidad para expresarse a través del sonido articulado o de otros sistemas de signos [1]. Éste depende del funcionamiento de múltiples zonas del cerebro, que se localizan principalmente en el hemisferio izquierdo, específicamente en la región perisilviana. El daño en estas regiones o en los circuitos que las relacionan ocasiona afasia, alteración que afecta tanto el lenguaje oral como escrito [2].

1.1.1. DEFINICIÓN DE AFASIA

La afasia se conoce como la pérdida o trastorno de la capacidad del habla debida a una disfunción en las áreas del lenguaje de la corteza cerebral [1]. En la inmensa mayoría de las personas diestras, en congruencia con la dominancia manual, y en el 70 % de las zurdas, en incongruencia con la dominancia manual, las lesiones que producen afasia se asientan en el hemisferio izquierdo. Además, dichas lesiones no solo afectan a las áreas corticales clásicas de Broca o de Wernicke, sino también a otras zonas como el área motora suplementaria o las conexiones entre ellas y los núcleos grises basales [3].

Se trata como un trastorno del lenguaje adquirido a consecuencia de un daño cerebral, que por lo general compromete alguna o todas sus modalidades: expresión y comprensión del lenguaje oral, escritura y comprensión de lectura. Cada una de éstas puede verse afectada cualitativa y cuantitativamente de forma diferente dependiendo de la topografía de la lesión cerebral, coexistiendo incluso con otras deficiencias en el procesamiento cognitivo como la anomia, dificultad para evocar las palabras [2].

1.1.2. CAUSAS DE LA AFASIA

La afasia se puede producir por una de las siguientes causas: enfermedad cerebrovascular o ictus, traumatismo craneoencefálico (TCE), tumor, procesos infecciosos o inflamatorios, y enfermedades neurodegenerativas progresivas como la Enfermedad de Alzheimer o la afasia progresiva primaria. Ésta última es un deterioro del lenguaje como consecuencia de un proceso neurodegenerativo que afecta fundamentalmente a las regiones frontal y temporal del hemisferio dominante [2].

ICTUS COMO CAUSA PRINCIPAL

Sin embargo, la causa más frecuente de afasia es el ictus. Se calcula que en España cerca del 80% de las personas con daño cerebral adquirido (DCA) han sufrido un infarto cerebral [4].

El ictus es una de las enfermedades neurológicas de mayor prevalencia y gravedad, con un impacto significativo en la salud pública. En España, cada año se registran entre 110.000 y 120.000 nuevos casos, y se estima que 1 de cada 4 personas lo sufrirá a lo largo de su vida. Esta afección no solo es la primera causa de discapacidad adquirida en adultos, sino también la segunda causa de muerte en el país, ocupando el primer lugar entre las mujeres. Su relevancia sanitaria se evidencia aún más al considerar que el ictus representa el 70% de los ingresos neurológicos en los hospitales [5].

Una de las consecuencias más devastadoras del ictus es la afasia. La afasia se produce entre el 15% y el 42% de los pacientes que sobreviven a un ictus, convirtiendo esta enfermedad en la causa más común de este trastorno. Se calcula que, a nivel global, la incidencia anual de afasia supera los 180.000 nuevos casos en países como Estados Unidos, afectando a un total de 2 millones de personas. En España, la cifra también es significativa, con miles de pacientes enfrentándose a esta condición cada año.

Comparada con otras enfermedades neurológicas como la Enfermedad de Parkinson, la parálisis cerebral o la distrofia muscular, la afasia es más frecuente, aunque paradójicamente mucho menos conocida. Este desconocimiento agrava el impacto emocional y social de quienes la padecen, al enfrentarse no solo a barreras comunicativas, sino también al aislamiento y la falta de comprensión por parte de su entorno [6].

Se conocen dos tipos de enfermedades cerebrovasculares, representadas visualmente en la Figura 1:

- **Ictus isquémico**: es causado por un coágulo que bloquea o interrumpe el riego sanguíneo en una parte del cerebro provocando la muerte del tejido neuronal por falta de oxígeno y nutrientes. Es el tipo más común, un 80% de los ataques cerebrales son isquémicos

- **Ictus hemorrágico**: es causado por la rotura de un vaso sanguíneo vertiendo la sangre dentro de la cavidad craneal. El hematoma resultante comprime y daña el tejido de una forma más destructiva [7].

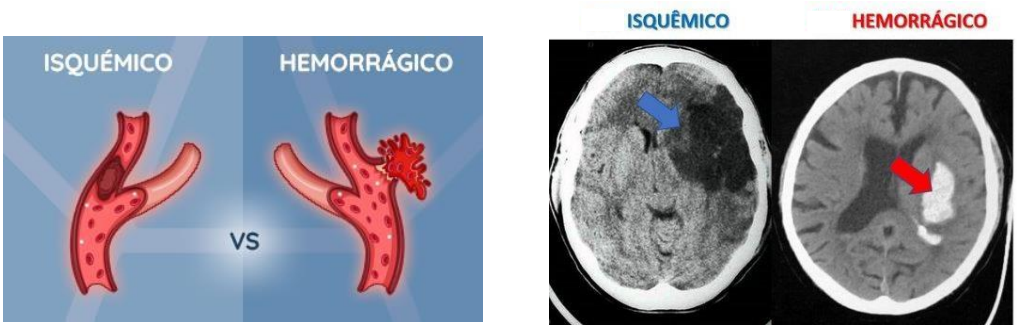

Figura 1. Ictus isquémico e ictus hemorrágico, diferencias fisiológicas representadas en un TC. En el ictus isquémico se distingue una zona de hipodensidad, mientras que en el hemorrágico se distingue una zona hiperdensa [4].

EVALUACIÓN DEL ICTUS Y DIAGNÓSTICO DE LA AFASIA

Cuando alguna de estas dos situaciones afecta a las regiones cerebrales encargadas del lenguaje, el paciente puede presentar afasia desde los primeros minutos del ictus [4].

La afasia se incluye como un parámetro en la escala de NIHSS (National Institutes of Health Stroke Scale), una herramienta esencial para medir la gravedad del daño neurológico. Esta escala analiza 11 aspectos clave, incluyendo movilidad, campo visual, funciones motoras y lenguaje, entre otros. El ítem de lenguaje resulta especialmente relevante para detectar la presencia y gravedad de la afasia, ya que evalúa la capacidad del paciente para comprender, nombrar y repetir palabras. Con una puntuación que oscila entre 0 y 42, donde un puntaje más alto indica mayores déficits, esta herramienta no solo permite identificar rápidamente la afasia, sino también cuantificar su impacto y orientar estrategias de rehabilitación lingüística en fases tempranas del tratamiento [8]. (ver anexo C.1 NIHSS)

Lo más común es que el médico que trata la lesión cerebral sea quien primero detecte la afasia. Habitualmente, se realiza una resonancia magnética o tomografía computarizada (TC) para confirmar la lesión cerebral y determinar su ubicación exacta. Además, el médico evalúa la capacidad de comprender y producir lenguaje mediante pruebas para verificar cómo el paciente sigue órdenes, responde preguntas, nombra objetos y mantiene una conversación [9].

Si se sospecha de afasia, generalmente se remite al paciente a un médico rehabilitador experto en foniatría, quien realiza un examen integral de las capacidades de comunicación. Este examen incluye

una evaluación detallada de la capacidad del paciente para hablar, expresar ideas, participar en conversaciones sociales, comprender el idioma, leer y escribir [9].

En definitiva, en la mayoría de los casos, la afasia sucede de manera repentina debido a un ictus. En casos menos comunes, como aquellos causados por un tumor, el desarrollo de la afasia es más lento. En estos casos, se suelen solicitar pruebas como una resonancia magnética o una TC para medir el grado de daño cerebral existente [7].

1.1.3. CLASIFICACIÓN DE LAS AFASIAS

A la hora de clasificar los tipos de afasia es importante tener en cuenta que están muy relacionados con la localización de la lesión. Según la zona afectada del cerebro, el paciente presenta distintos problemas de comunicación.

NEUROANATOMÍA DEL LENGUAJE

Rescatando el estudio de la afasia de Wernicke y Lichtheim, Norman Geschwind propuso en 1965 el modelo neoconexionista [4]. Este autor desarrolló un modelo más completo del procesamiento del lenguaje en el cerebro profundizando en la idea de que el cerebro era un conjunto de partes conectadas en las que cada proceso ocurría de forma secuencial. En el circuito del lenguaje, conocido como red perisilviana del lenguaje, identificó los siguientes elementos esenciales (ver Figura 2):

- **Área de Broca**: Encargada de la expresión, coordina los movimientos articulatorios para la producción del habla.
- **Área de Wernicke**: Encargada de la comprensión, transforma la información auditiva de las palabras en unidades de significado.
- **Área Conceptual**: Un área de asociación clave que implica la conexión de regiones auditivas, visuales y somestésicas (sensibilidad: tacto, presión, dolor…). Lichtheim no pudo determinar su localización, fue Geschwind quien consiguió situarla en la circunvolución angular y supramarginal.
- **Fascículo arqueado**: Encargado de la repetición, conecta las áreas de Broca y Wernicke.
- **Giro angular**: Encargado de transformar el modelo visual de una palabra, es decir la lectura, a su forma auditiva.

El procesamiento secuencial del lenguaje se daría de la siguiente forma: al oír una palabra, la información viajaría hasta el Área de Wernicke, donde se comprendería; si quisiéramos repetirla en voz alta, la información pasaría a través del fascículo arqueado hasta el Área de Broca, donde se crearía la forma sonora de esa palabra; de ahí, iría al área motora primaria que controla el movimiento de los músculos fonatorios. En cambio, si se leyese una palabra, la información llegaría desde las áreas visuales y de aquí al giro angular, que transmitiría la información al área de Wernicke [4].

En la Figura 2, se representa gráficamente la diferencia entre el circuito secuencial del lenguaje al escuchar una palabra y al leerla escrita. Mientras que al escuchar una palabra el circuito comienza en el área auditiva primaria, al leerla, el proceso se inicia en el área visual primaria ya que la información se obtiene a través de la vista.

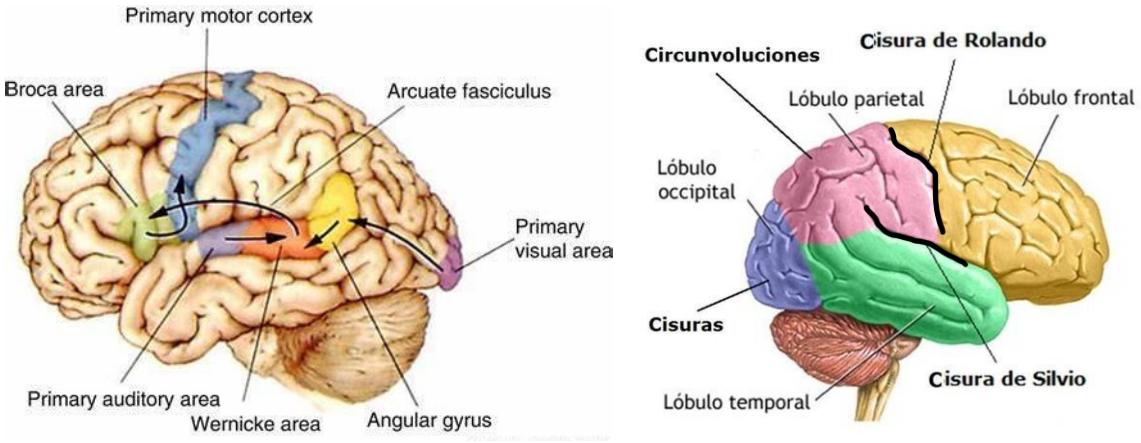

Figura 2. Representación gráfica del procesamiento secuencial del lenguaje, tanto visual como auditivo, en las distintas áreas cerebrales [10].

Este modelo tuvo un impacto significativo en la neurolingüística al proporcionar explicación para entender diversos trastornos del lenguaje, como la afasia, basándose en la interrupción de diferentes partes del circuito. Sin embargo, este modelo soporta críticas al ser considerado demasiado simple en comparación con la complejidad real del procesamiento del lenguaje, no considerar suficientemente la plasticidad cerebral y subestimar el papel del hemisferio derecho en el procesamiento del lenguaje [11]. Esto último hace referencia a la afasia cruzada, clasificada como afasia especial, que podemos encontrar en un sujeto diestro por una lesión en el hemisferio derecho. La incidencia de esta afasia es de un 2% a un 5%. Un 70% de estos pacientes presenta un patrón en espejo, es decir, clínicamente los mismos cuadros sindromáticos observados por lesiones del hemisferio izquierdo. El otro 30% muestra un patrón atípico, sin correlación entre el sitio de la lesión y el perfil clínico esperable [2].

A pesar de sus limitaciones, el modelo de Wernicke-Geschwind sigue siendo un punto de partida importante para la neuroanatomía del lenguaje inspirando modelos más sofisticados de redes cerebrales del lenguaje [11]. Actualmente gracias a investigaciones con herramientas como la Resonancia Magnética o la TC, se ha confirmado que los modelos conexionistas estaban acertados en líneas generales. No obstante, estos estudios también han aportado nuevas evidencias que matizan y enriquecen los modelos clásicos.

Por ejemplo, se ha demostrado que para que se produzca una afasia de Broca, las lesiones han de extenderse más allá de la propia área de Broca, es decir, un daño exclusivamente en esa zona provocaría dificultades en el habla, pero sin sufrir una alteración lingüística significativa. Además, se ha descubierto que esta área, encargada de la producción del habla, no se activaría a la hora de leer en voz alta. Todos estos datos han llevado a formular nuevas propuestas sobre la organización cerebral del lenguaje donde se incluyen nuevas áreas como las corticales y subcorticales a un complejo sistema neuronal entendido como una formación de circuitos que conectan múltiples áreas [4].

TIPOS DE AFASIAS

De esta manera, siguiendo el modelo de Wernicke-Geschwind existen tres variables que ayudan a un profesional a clasificar el tipo de afasia concreta [2].

- Fluidez: cantidad de palabras que emite el afectado al expresar una oración.
- Comprensión: capacidad del afectado para entender palabras, frases y textos.
- Repetición: capacidad del afectado para replicar sonidos o palabras.

La primera clasificación se da entre dos grandes grupos, fluentes y no fluentes, cuyas características se pueden observar de manera esquemática en la Tabla 1.

Las afasias fluentes se caracterizan por un discurso productivo, en cuanto a la cantidad de palabras. El discurso es poco informativo, presentando más palabras funcionales que de contenido. La articulación suele ser sin esfuerzo y adecuada, así como la longitud del enunciado y la línea melódica. En cuanto a la gramática, se encuentra generalmente conservada, ya que pueden presentar paragramatismo. Los errores más comunes son: las parafasias fonémicas, semánticas, verbales y neológicas. Estas afecciones se producen por lesiones localizadas detrás de la cisura de Rolando (ver Figura 2).

Las afasias no fluentes, en la mayoría de los casos, presentan reducción del discurso tanto cualitativa como cuantitativamente. Se caracterizan principalmente por la dificultad para iniciar enunciados. La articulación es laboriosa y en los cuadros más graves, se puede observar producción de sílabas aisladas y estereotipias, llegando en algunos casos al mutismo. La longitud de los enunciados es breve y la prosodia se encuentra alterada. La morfosintaxis está alterada, observándose más palabras de contenido que funcionales. Además, poseen dificultad para acceder a los verbos con carga semántica. Ésta puede coexistir con disartria y apraxia del habla. Las lesiones que producen estos tipos de afasias se localizan por delante de la cisura de Rolando.

Discurso de tipo fluente	Discurso de tipo no fluente
Inician los enunciados sin dificultad	Dificultad para iniciar los enunciados
Gran cantidad de palabras por minuto	Pocas palabras por minuto
Poco informativo	Informativo
Más palabras funcionales que de contenido	Más palabras de contenido que funcionales
Articulación sin esfuerzo y adecuada	Articulación alterada
Longitud del enunciado conservada	Enunciados de pocas palabas
Línea melódica adecuada	Línea melódica alterada
Frecuentes parafasias	Pocas parafasias
Paragramatismo	Agramatismo
Lesión por detrás de la cisura de Rolando	Lesión por delante de la cisura de Rolando

Tabla 1. Características clínicas del tipo de discurso en relación con la fluidez [2].

A continuación, se debe comprobar si está conservada la comprensión auditiva y finalmente, establecer si es capaz o no de repetir enunciados (ver Figura 3).

AFASIAS FLUENTES

Afasia de Wernicke (sensitiva): *(FLUENTE, NO COMPRENDE, NO REPITE)*

Se caracteriza por un discurso fluido, pero poco informativo, a menudo con logorrea y parafasias (errores de palabras). La comprensión auditiva y la repetición están gravemente afectadas, y los pacientes generalmente no tienen conciencia de sus errores (anosognosia). Esto dificulta significativamente la comunicación funcional y suele requerir estrategias visuales o contextuales para mejorar la interacción.

Afasia transcortical sensorial: (FLUENTE, NO COMPRENDE, REPITE)

Incluye un discurso fluente con parafasias graves y ecolalia. Los pacientes pueden repetir estímulos verbales sin comprenderlos. Las dificultades de denominación y comprensión son significativas, con alteraciones asociadas en la lectura y escritura. A menudo, las lesiones se localizan en áreas alrededor de la región de Wernicke, pero sin afectarla directamente.

Afasia de conducción: (FLUENTE, COMPRENDE, NO REPITE)

Se caracteriza por un discurso fluente con parafasias fonémicas y autocomprobaciones frecuentes. Aunque la comprensión es relativamente buena, los pacientes tienen dificultades marcadas en la

repetición y en tareas que implican ensamblaje fonológico. Esto se relaciona con lesiones en el fascículo arqueado, que conecta las áreas de Broca y Wernicke.

Afasia anómica: (FLUENTE, COMPRENDE, REPITE)

La principal dificultad radica en encontrar palabras (anomia), lo que lleva a pausas y circunloquios en el discurso. La comprensión, la repetición y otras habilidades lingüísticas están preservadas, excepto en tareas de alta complejidad. Aunque es menos grave, esta afasia puede afectar considerablemente la vida diaria al dificultar la comunicación espontánea.

AFASIAS NO FLUENTES

Afasia global: (NO FLUENTE, NO COMPRENDE, NO REPITE)

Esta es la forma más grave de afasia, con pérdida significativa de todas las funciones lingüísticas. El discurso está limitado a estereotipias o mutismo, con comprensión y escritura completamente alteradas. Las lesiones suelen abarcar toda la región perisilviana. La rehabilitación en estos casos se enfoca en mejorar aspectos básicos de la comunicación funcional y la calidad de vida.

Afasia transcortical mixta: (NO FLUENTE, NO COMPRENDE, REPITE)

Combina características de la afasia global, pero con conservación de la repetición. Los pacientes presentan ecolalia y una producción verbal extremadamente limitada, con graves alteraciones en la comprensión y la escritura. Las lesiones suelen ubicarse en regiones más externas, como las áreas asociativas, preservando los circuitos básicos de repetición.

Afasia de Broca (motora): (NO FLUENTE, COMPRENDE, NO REPITE)

El discurso es no fluente y laborioso, con frases cortas y agramáticas. La comprensión está relativamente preservada para enunciados simples, pero la repetición está gravemente afectada. Puede asociarse a apraxia del habla y disartria. Aunque la rehabilitación puede ser efectiva, el progreso es lento debido a las dificultades motoras asociadas.

Afasia transcortical motora: (NO FLUENTE, COMPRENDE, REPITE)

El discurso es no fluente y breve, pero la repetición está preservada. Los pacientes muestran ecolalia y dificultad para iniciar el habla. La comprensión está mayormente conservada, pero la escritura puede estar afectada debido a problemas de iniciación. Este tipo de afasia está relacionado con lesiones en áreas frontales superiores alejadas de la región de Broca.

[2] [4] [6]

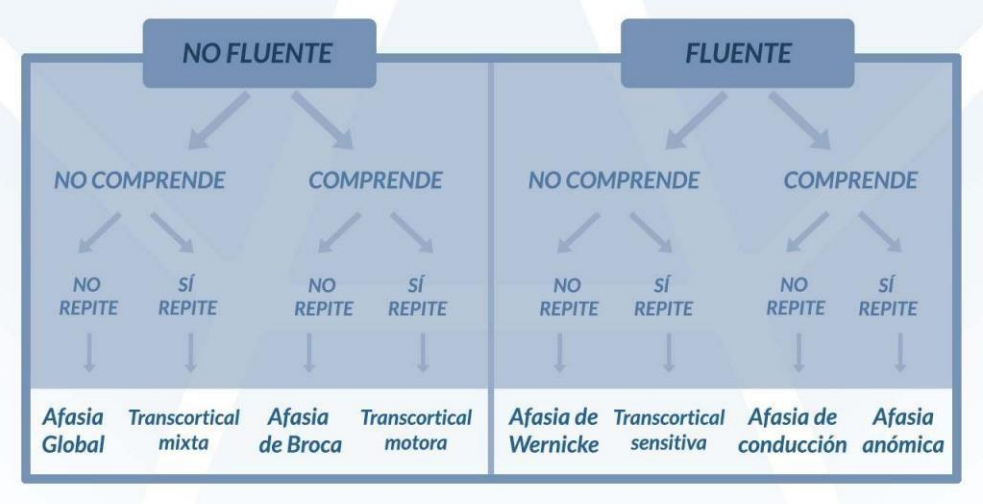

Figura 3. Algoritmo para la clasificación clínica de las afasias [4].

Existen posibles comorbilidades de las afasias, es decir, presencia de otros trastornos asociadas a esta condición. Por ejemplo [12]:

- Alexia y agrafia: Son alteraciones variables de la expresión y comprensión lectora y escrita.
- Trastornos cognitivos: En la memoria, atención, funciones ejecutivas, etc. Suele producir dificultades para seguir instrucciones, retener y almacenar la información, bradipsiquia y otro tipo de alteraciones que afectan a la expresión y comprensión verbal.
- Trastornos motores:
  - Hemiparesia o hemiplejia: en las extremidades.
  - Disartria y apraxia del habla: en el habla.
  - Apraxias ideomotoras e ideatorias: en el control motor voluntario.

- Trastornos sensoriales:
  - Diplopía y hemianopsia: visuales.
  - Hipoacusias: auditivos.
  - Hipoestesia: sensibilidad somatoestésica.

- Alteraciones neuropsiquiátricas: Como depresión, ansiedad, apatía, labilidad emocional, etc.
- Disfagia orofaríngea: Alteraciones deglutorias.

1.1.4. LIMITACIONES EN LA VIDA COTIDIANA

El lenguaje distingue a los humanos de otras especies. Lo usamos constantemente para decirnos unos a otros lo que sentimos, pensamos y necesitamos. Perder la capacidad de comunicarse supone un fuerte impacto en la vida cotidiana de las personas que padecen de afasia [10]. Su rol en la sociedad cambia por completo y, al no poder comunicarse, pasan a sentirse personas dependientes. Este cambio repentino en sus vidas puede afectar gravemente a nivel emocional, pudiendo desarrollar depresión transitoria. Es muy importante prestar el cuidado que se merece, si no, las consecuencias pueden llegar a ser devastadoras: tendencia al aislamiento, falta de control de los impulsos, riesgo de padecer otras enfermedades mentales, abuso de sustancias nocivas, debilitamiento del sistema inmunológico o tendencia a la autolesión o al suicidio en los casos más extremos [4].

Aunque una persona con afasia puede tener dificultad para acceder a la ideas y pensamientos a través del lenguaje, su inteligencia permanece prácticamente intacta. Sin embargo, dado que la mayoría de los trabajos requieren habilidades de habla y lenguaje, la afasia puede dificultar ciertos tipos de empleo haciendo que muchas personas tengan que cambiar o incluso dejar de trabajar. Esto supone otro cambio muy grande para ellos al que hacer frente [6].

Numerosos estudios han demostrado que las personas con afasia enfrentan limitaciones significativas en su vida cotidiana, afectando su integración social, su participación comunitaria y su calidad de vida. Estas limitaciones abarcan desde barreras en las relaciones personales hasta dificultades para mantener actividades productivas, lo que las sitúa en una posición de mayor vulnerabilidad social y emocional.

En el estudio “Community Integration and Quality of Life in Aphasia after Stroke” publicado en ResearchGate [13], en el que participaron 30 personas con afasia y 42 controles pareados por edad y nivel educativo, se evaluaron variables como el estado socioeconómico, movilidad y actividades de la vida diaria (Índice de Barthel Modificado), función lingüística (Test de Cribado de Afasia de Frenchay), depresión (Escala de Depresión Geriátrica), integración comunitaria (Cuestionario de Integración Comunitaria) y calidad de vida (Escala de Calidad de Vida en Ictus y Afasia-39) (ver anexo C.2 CIQ). Los resultados obtenidos en el estudio se pueden observar en la Figura 4 y Figura 5.

Se observó que las personas con afasia tienen una integración social y productiva significativamente menor en comparación con grupos de control. Las personas con afasia no solo pasaron menos tiempo fuera del hogar, sino que también tuvieron un menor número de contactos sociales y participaron menos en actividades comunitarias (8.5 frente a 18.3 puntos en la puntuación del Cuestionario de Integración

Comunitaria). Estas limitaciones afectaron directamente su calidad de vida, la cual se correlacionó con factores como la movilidad, las actividades diarias y el estado emocional, siendo la depresión el principal factor predictivo de una baja calidad de vida [13]. Incluso cuando sus habilidades físicas son comparables, las personas con afasia tienen una calidad de vida significativamente inferior a la de personas sin esta condición. Las barreras comunicativas que enfrentan no solo impactan en su funcionalidad diaria, sino que también incrementan su vulnerabilidad emocional, con altas tasas de ansiedad y depresión [14]. En comparación con personas sin afasia, hasta el 93% de las personas con afasia en edad laboral mostraron signos de estrés emocional y angustia psicológica grave, frente al 50% en quienes no presentan esta condición. Estas conclusiones subrayan la necesidad de abordar tanto los desafíos lingüísticos como los aspectos emocionales y sociales en los programas de rehabilitación [15].

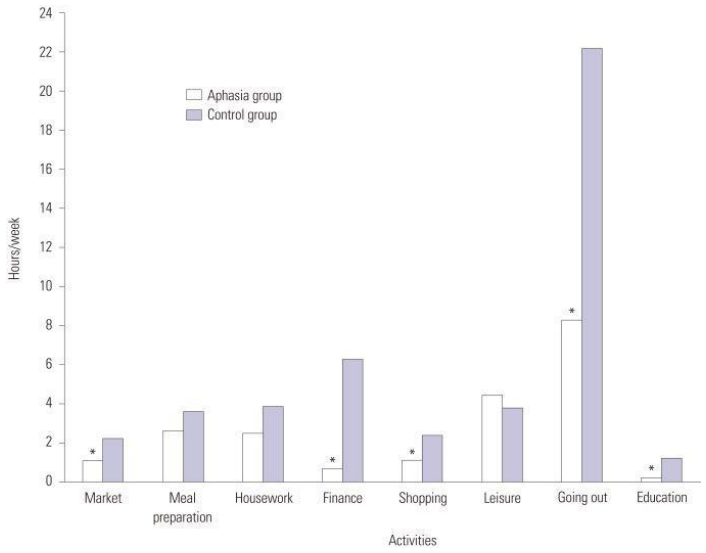

Figura 4. Comparación del tiempo dedicado a realizar las subunidades del Cuestionario de Integración Comunitaria entre los grupos con afasia y el grupo de control. La mayoría de las actividades fuera del hogar se redujeron significativamente en el grupo con afasia (*p<0.05) [13].

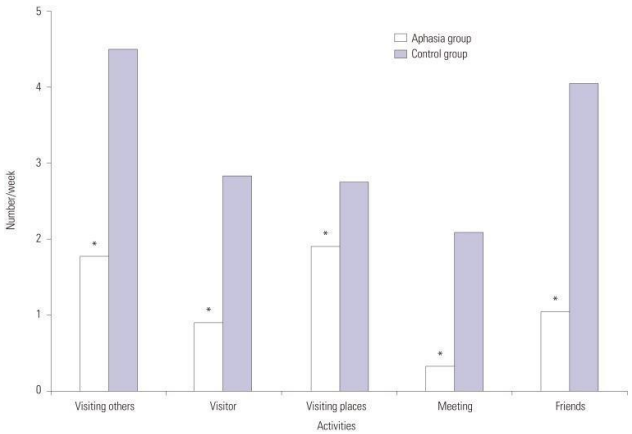

Figura 5. Número promedio de actividades realizadas durante una semana en los grupos con afasia y de control. El número de contactos sociales se redujo significativamente en el grupo con afasia (*p=0.00) [13].

1.2. OBJETIVOS

Este Trabajo de Fin de Grado tiene como objetivo el diseño e implementación de un sistema de generación de vídeos mediante Inteligencia Artificial (IA) Generativa como técnica emergente en la rehabilitación de personas con afasia.

De esta manera, se podrán recrear situaciones de la vida cotidiana que supongan una dificultad para los pacientes, de forma que puedan practicar de manera visual, contextual y adaptada a sus necesidades específicas.

Para alcanzar este propósito:

1. Se evaluarán las distintas herramientas actuales de IA generativa que permitan la creación de vídeos con características esenciales como sincronización de labios, cambios de escenario y emociones relevantes para cada contexto.
2. Se propondrá la combinación de herramientas óptima para la generación de audios, escenarios y avatares.
3. Se detallarán posibles dificultades y limitaciones técnicas a la hora de implementar el sistema en entornos clínicos como hospitales.
4. Se incluirán propuestas para mejoras y continuaciones en trabajos de fin de grado posteriores.

Este Trabajo de Fin de Grado surge de la colaboración del Laboratorio de Robótica y Control (Robolabo) la Escuela Técnica Superior de Ingenieros de Telecomunicación de la Universidad Politécnica de Madrid y el servicio de Neurología del Hospital Universitario La Paz.

1.3. ORGANIZACIÓN DEL DOCUMENTO

El documento consta de la siguiente estructura:

- **Introducción**: Se realiza una introducción a la afasia, determinando sus causas y exponiendo las limitaciones cotidianas que desencadena. Asimismo, se explican los objetivos del presente trabajo y la organización de este.

- **Estado del arte**: Se desarrolla el estado del arte desde la perspectiva clínica y desde la perspectiva tecnológica.

- **Métodos**: Se detalla la metodología seguida para el diseño e implementación del sistema de generación de vídeos mediante IA generativa. Se explican la propuesta de herramientas a integrar, requerimientos computacionales necesarios para llevar a cabo el proyecto y la arquitectura del sistema.

- **Resultados**: Se exponen los resultados obtenidos.

- **Conclusiones**: Se muestran las conclusiones a las que se ha llegado durante la realización del trabajo y se proponen las líneas de investigación futuras.

2. ESTADO DEL ARTE

2.1. TÉCNICAS DE REHABILITACIÓN

El tratamiento de la afasia se encuentra integrado dentro de unidades multidisciplinares de neurorrehabilitación donde trabajan los diferentes profesionales del ámbito sanitario como médicos rehabilitadores y foniatras, logopedas, fisioterapeutas, terapeutas ocupacionales o neuropsicológicos.

Los logopedas son los implicados directamente en la prevención, la evaluación, el diagnóstico y la rehabilitación de los problemas comunicativos que afectan al lenguaje, como ocurre con la afasia, pero también al habla (disartria y apraxia del habla), la voz (disfonías) o de la deglución (disfagia orofaríngea), entre otros aspectos.

La participación de neuropsicólogos es crucial a la hora de la evaluación y tratamiento de las funciones cognitivas afectadas por lesiones cerebrales, contribuyendo a la recuperación de habilidades como la atención, memoria, funciones ejecutivas, habilidades sociales y comportamiento emocional, que son fundamentales para una comunicación efectiva [12].

COMIENZO DE LA REHABILITACIÓN

La rehabilitación de la afasia debe iniciarse lo antes posible, cuando lo permita el estado general de salud del paciente. Los estudios sobre la recuperación de las afasias han determinado que las mejoras más importantes se producen en los primeros dos o tres meses, alcanzando los mejores resultados tras el año de intervención [16]. La recuperación espontánea se registra en los seis primeros meses tras el episodio, mientras que sólo se esperan mejoras espontáneas mínimas de las funciones del lenguaje pasado el año. Por ello, es importante llevar a cabo una intervención temprana, para obtener mejores resultados y un mejor pronóstico.

Se ha demostrado que la rehabilitación de la afasia es eficaz cuando se administra de forma intensiva y/o durante periodos prolongados. Estudios recientes revelaron que la terapia intensiva del habla y el lenguaje, administrada durante 3 horas al día durante dos semanas, puede inducir mejoras significativas en las funciones del lenguaje en pacientes con afasia crónica. Estos cambios se correlacionaron con una reorganización en las áreas perilesionales del cerebro, destacando la importancia de la plasticidad cerebral en la recuperación tras un ictus [17] [18]. La ley de Hebb [19] puede ilustrar cómo funciona la plasticidad cerebral: cuando un circuito neuronal es activado de forma frecuente y sincronizada, la mayoría de sus conexiones se verán reforzadas; de forma opuesta, aquellas neuronas que “disparan” fuera del circuito se verán sometidas a un debilitamiento sináptico y, por tanto, a una pérdida conectiva.

Partiendo entonces de esta teoría, una terapia de habla y lenguaje intensiva obtendrá mejores resultados, ya que, por cada uso de una palabra en el contexto apropiado, se podrá esperar un refuerzo proporcional de las sinapsis de circuitos relevantes para el habla y el lenguaje, especialmente si la terapia es administrada en alta frecuencia. Además, dado que existen conexiones funcionales entre las áreas sensoriomotoras y el córtex lingüístico, aquellas terapias que tengan por objetivo relacionar lenguaje y acción conducirán a una mejora más eficaz de los déficits asociados a la afasia [19].

2.1.1. TÉCNICAS ACTUALES

Existen numerosas técnicas para la rehabilitación de la afasia. La terapia específica, al igual que la duración e intensidad de esta, dependen del tipo de pérdida del lenguaje que tenga la persona [7]. Sin embargo, las guías internacionales actuales recomiendan la terapia del habla y lenguaje (SLT, por sus siglas en inglés) para los supervivientes de un ictus [20].

TERAPIA DEL HABLA Y DEL LENGUAJE: LOGOPEDIA

Existen artículos donde se realiza una revisión sistemática que evalúa la eficacia de la terapia del habla y lenguaje en personas con afasia tras una enfermedad cerebrovascular comparando la SLT con la ausencia de terapia, apoyo social o diferentes enfoques de SLT.

Los resultados indican que la SLT proporciona mejoras significativas en la comunicación funcional, así como en las habilidades de lectura, escritura y lenguaje expresivo, en comparación con la no intervención. La revisión también analizó diferentes modalidades de SLT, encontrando que terapias de alta intensidad, mayor dosis o duración ofrecían beneficios adicionales en la comunicación funcional. No obstante, estas modalidades presentaron tasas más altas de abandono, lo que sugiere que no todos los pacientes las toleran bien, recalcando la necesidad de personalizar la intensidad y duración de la terapia según las necesidades individuales de cada paciente [20].

El tratamiento puede ser individual o en grupo con un logopeda o, en ocasiones, utilizando un ordenador. Puede incluir ejercicios de lectura, escritura, seguir instrucciones y repetir lo que dice el terapeuta [7].

Ejercicios sin apoyo visual

La mayoría de las actividades que se suelen practicar consisten en ejercicios sin apoyo visual (ver Tabla 2). Estas técnicas ayudan al paciente a la hora de retener la información en su cabeza e imaginarse el ejercicio que le están proponiendo.

Tipo de Actividad	Descripción	Objetivos
Nombrar e Identificar	Los participantes observan objetos o imágenes y se les pide nombrarlos o identificarlos.	Mejora la recuperación de palabras y amplía el vocabulario.
Seguir Instrucciones	Las actividades consisten en seguir instrucciones verbales o escritas tanto simples como complejas.	Mejora la comprensión de instrucciones verbales.
Tareas de Secuenciación	Se pide a los participantes organizar los pasos de una actividad común (por ejemplo, preparar té) en el orden correcto.	Ayuda a planificar y ejecutar tareas de forma lógica.
Practicar Preguntas	Practica responder preguntas de sí/no y preguntas más complejas como ‘quién, qué, dónde, cuándo, por qué’.	Mejora la toma de decisiones y la capacidad de expresar preferencias.
Completar Refranes	Se proporciona la primera parte de un refrán y el participante debe completarlo.	Fomenta la recuperación de información almacenada y mejora la fluidez verbal.
Palabras Encadenadas	Cada palabra debe comenzar con la última sílaba de la palabra anterior.	Mejora la asociación de ideas y la capacidad de iniciar palabras.
Memorización de Listas de Objetos	Se presenta una lista de objetos que el participante debe recordar.	Ejercita la memoria inmediata y la capacidad de relacionar conceptos.

Tabla 2. Ejemplos de ejercicios prácticos de logopedia para rehabilitación de afasias [21].

Ejercicios con apoyo visual

De vez en cuando en SLT se incluyen ejercicios con dibujos impresos o en el ordenador. Por ejemplo, hacer la misma actividad de retener una serie de objetos, pero enseñándoles previamente estos objetos en forma de dibujos. También se puede practicar la narrativa de historias. Los terapeutas les proponen un personaje con unas características específicas y ellos mismos deben crear la historia en base a lo que ven en el dibujo. La implementación de recursos visuales, como pictogramas y fotografías, en cuadernos de apoyo a la comunicación, refuerza el mensaje hablado y facilita la comprensión, especialmente cuando se señalan elementos clave durante la interacción [22]. Sin embargo, los logopedas suelen tener bastantes dificultades para encontrar elementos visuales de interés que usar en la rehabilitación de estos pacientes, los cuales son personas mayores normalmente. Muchas modalidades de SLT utilizan materiales infantiles, lo que limita su aplicabilidad en la vida cotidiana de los adultos. Este es uno de los principales desafíos de esta terapia, ya que genera altas tasas de abandono temprano, de hasta el 30%, debido a que los pacientes sienten que sus preferencias no están siendo satisfechas [20].

OTRAS TERAPIAS

Al margen de la logopedia se están probando otro tipo de terapias de rehabilitación como las técnicas de estimulación cerebral no invasiva donde se encuentran la Estimulación Magnética Transcraneal (EMT) y la Estimulación Transcraneal de Corriente Directa (ETCD). Estas técnicas buscan modular la actividad cerebral para potenciar la recuperación del lenguaje. Por ejemplo, la EMT se ha empleado para estimular áreas perilesionales del hemisferio izquierdo, promoviendo la plasticidad cerebral y facilitando la reorganización funcional del cerebro [23] [24].

Como rehabilitaciones complementarias existen:

- **Terapia inducida por restricción**: terapia física de menor duración que todavía no está cubierta por el seguro médico y que de utilizarse, sería de complemento a la rehabilitación habitual. Se basa en la idea de que un paciente con afasia será propenso a utilizar un lenguaje telegráfico para comunicarse o incluso evitará cualquier tipo de comunicación verbal y se limitará a comunicarse por gestos. La terapia consiste en forzarles a utilizar el lenguaje hablado y evitar estos aprendizajes estratégicos que hacen los pacientes para eludir sus dificultades [19] [25].
- **Terapia de entonación melódica**: utiliza el poder de la música como rehabilitación en pacientes con afasia no fluente y buena comprensión. Los pacientes son capaces de producir oraciones en base a una melodía artificial. Sin embargo, se ha observado que hay pacientes que hablan mejor con la entonación melódica, pero no logran generalizar lo aprendido a una conversación natural [4] [21].

2.1.2. NUEVAS TÉCNICAS BASADAS EN DOBLAJE

Numerosos estudios han demostrado la eficacia de nuevas técnicas de rehabilitación como complemento a las terapias tradicionales para maximizar la rehabilitación del lenguaje. Por ejemplo, se probó un sistema de rehabilitación del habla asistido por computadora con generación de vídeos espejo [26].

Este sistema combina tecnologías de inteligencia artificial con la terapia espejo para ayudar a pacientes con afasia a mejorar sus habilidades lingüísticas y comunicativas.

La terapia espejo es una técnica de rehabilitación utilizada para aliviar el dolor y mejorar el movimiento en personas con diversas condiciones neurológicas, especialmente aquellas que han sufrido lesiones en

las extremidades o amputaciones. El principio fundamental de la terapia espejo se basa en el concepto de las neuronas espejo en el cerebro. Estas neuronas son células cerebrales que se activan tanto cuando realizamos una acción particular como cuando observamos a alguien más realizando la misma acción. Al usar espejos para crear ilusiones visuales, el cerebro es engañado para percibir el reflejo de la extremidad no afectada como si fuera la afectada. Esta retroalimentación visual puede generar cambios significativos en la percepción y el control motor del cerebro. Este estudio se basa en la idea de explorar la terapia espejo como un tratamiento para la afasia. Se generan vídeos con el rostro y la voz del paciente utilizando deep learning, lo que proporciona una retroalimentación visual y auditiva personalizada permitiendo que los pacientes se imiten a sí mismos en el vídeo espejo. Se incluyen ejercicios de palabras y formación de oraciones, donde los pacientes imitan los vídeos espejo, fortaleciendo la pronunciación y la estructura gramatical. Como resultado, se observaron mejoras significativas en ejercicios de vocales, palabras y oraciones, con tasas de éxito promedio del 83.9%, 74.3% y 77.8% respectivamente, demostrando ser una prometedora futura línea de investigación [26].

Otro estudio anterior ya abordó una mejora inesperada en la comunicación de un paciente con afasia de Broca tras recibir terapia espejo (MT) para la extremidad superior afectada por un ictus [27]. Inicialmente, no tenía capacidad de habla espontánea, escritura ni denominación de objetos. Su comunicación se limitaba a gestos. Sin embargo, la terapia espejo dirigida a la extremidad superior no solo mejoró la función motora, sino que también activó el sistema de neuronas espejo del área de Broca, favoreciendo la recuperación del habla sin intervención logopédica directa. Como resultado, se observó un incremento notable en el Índice de Eficiencia Comunicativa (CETI, por sus siglas en inglés) de 18/100 a 79/100, mejorando en áreas como expresión de emociones, conversación uno a uno, denominación de objetos, conversación espontánea y fluidez [27].

PROYECTO DULCINEA

En la línea de estos estudios basados en la terapia de espejo, surge DULCINEA, un proyecto innovador cuyo nombre hace referencia a: DUbbing Language-therapy CINEma-based in Aphasia post-stroke [28].

Surge de la colaboración entre el Hospital Universitario La Paz, la Universidad de Comillas y la asociación de pacientes, Afasia Activa, entre otros centros.

El objetivo principal e inicial era crear y validar una terapia basada en doblaje para mejorar la comunicación funcional en pacientes con afasia no fluente tras un ictus isquémico en el hemisferio izquierdo, sin evidencia en neuroimagen de lesiones en el hemisferio derecho. Otros de los criterios de inclusión en el proyecto son: haber completado previamente un programa estándar de terapia convencional del habla, tener lenguaje gravemente restringido, con mala repetición incluso de palabras simples y comprensión del lenguaje moderadamente preservada (es decir, puntuaciones que no superen el percentil 70 en la Escala de Repetición, además de superar el percentil 15 en la Escala de Comprensión Auditiva, como promedio de las subescalas de comprensión de palabras, órdenes y material ideacional complejo en el Examen Boston de Diagnóstico de Afasia (BDAE)) (ver anexo C.3 BDAE).

Como criterios de exclusión se destaca la participación simultánea en cualquier otro ensayo terapéutico que evalúe la recuperación del ictus o tener una condición clínica u otra característica que impida un seguimiento adecuado.

El programa consta de 16 sesiones distribuidas en 8 semanas, donde los pacientes practican el doblaje de escenas de series y películas mediante el software DULCINEA. Este sistema integra herramientas como zoom labial, sincronización audiovisual y ajuste de velocidad para trabajar aspectos clave del lenguaje: articulación, entonación, contexto emocional y precisión temporal. Las tareas se adaptan progresivamente a la dificultad y necesidades de cada paciente, comenzando con palabras simples y avanzando hacia frases más complejas. Además, los ejercicios pueden reforzarse en casa con la supervisión de un familiar.

Este estudio utilizó un diseño cruzado, representado gráficamente en la Figura 6, con dos secuencias de tratamiento (tratamiento-periodo de lavado/lista de espera-tratamiento) y dos fases (Fase 1 y Fase 2). Los pacientes fueron asignados aleatoriamente a uno de los siguientes grupos mediante un generador de números aleatorios informatizado proporcionado por un estadístico independiente.

- Grupo 1: La terapia comienza dentro de los primeros 3 meses tras la inclusión de los pacientes en el estudio, seguida de un periodo posterior de 3 meses sin terapia (periodo de lavado), sirviendo como grupo de control para la segunda fase del estudio.
- Grupo 2: La terapia comienza entre 3 y 6 meses después de la inclusión en el estudio. Durante los primeros 3 meses no reciben terapia del lenguaje (grupo de control en lista de espera) y actúan como el grupo de intervención activa en la segunda fase del estudio.

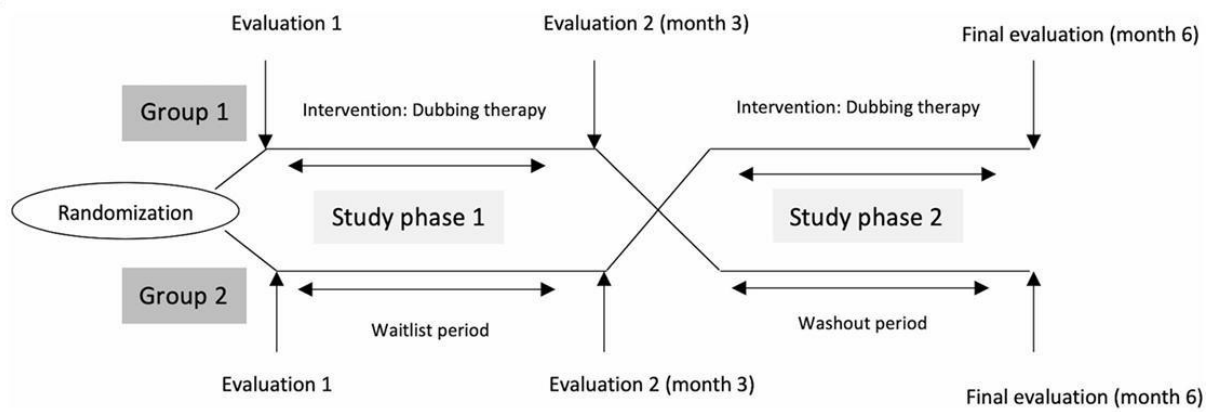

Figura 6. Diseño del estudio [28].

Este ensayo clínico mostró que el enfoque es factible y altamente aceptado por los pacientes, con una tasa de adherencia del 95.3%. Aunque no se encontraron diferencias estadísticamente significativas en los resultados de las evaluaciones estandarizadas (CAL, BDAE), el 85% de los pacientes reportaron mejoras subjetivas en sus habilidades comunicativas [29] [28].

A raíz del proyecto DULCINEA, en el que se ha utilizado material audiovisual de la serie *Cuéntame* vinculada a RTVE para desarrollar esta terapia, surge la necesidad de superar las limitaciones asociadas a la dependencia de contenidos preexistentes. De las numerosas temporadas de la serie, solo un número limitado de clips ha resultado útil para los objetivos terapéuticos. Para abordar esta limitación, se ha planteado la creación de vídeos generados mediante IA, simulando acciones cotidianas con las que los pacientes puedan practicar en un entorno clínico a modo de rehabilitación complementaria a las actividades habituales, ampliando las posibilidades terapéuticas sin depender de material externo.

2.2. MODELOS DE INTELIGENCIA ARTIFICIAL GENERATIVAS

La IA es una rama de la informática que busca crear sistemas capaces de realizar tareas que, cuando son realizadas por seres humanos, requieren de inteligencia. Estas tareas incluyen el reconocimiento de voz, la toma de decisiones, la percepción visual y la traducción de idiomas, entre otras [30].

Dentro de la IA distinguimos el Machine Learning y el Deep Learning (ver Figura 7).

- Machine Learning (ML) o aprendizaje automático: Es una subdisciplina de la IA que permite a las máquinas aprender de datos y mejorar su desempeño en tareas específicas de manera automática sin ser programadas explícitamente para ello. Utiliza algoritmos que identifican patrones en los datos y hacen predicciones basadas en ellos.

- Deep Learning (DL) o aprendizaje profundo: Es una subcategoría o tipo concreto del ML que emplea redes neuronales artificiales con múltiples capas que están jerarquizadas de forma que todas las neuronas se conectan con las de la capa siguiente, para modelar representaciones de datos complejas, emulando el aprendizaje humano. Este enfoque ha sido fundamental en avances recientes de la IA, permitiendo mejoras significativas en áreas como el reconocimiento de voz e imágenes [30] [31].

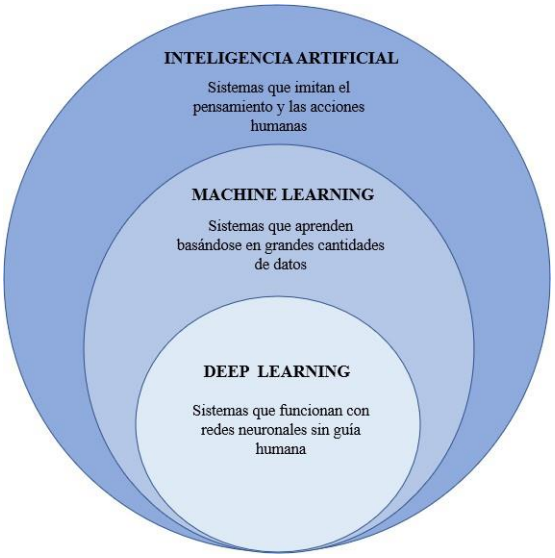

Figura 7. Representación gráfica de la relación entre IA, ML y DL [31].

DEFINICIÓN DE IA GENERATIVA

La IA Generativa es un tipo de DL diseñada para crear contenido audiovisual nuevo y original, como texto, imágenes, música o vídeos, a partir de datos existentes. A diferencia de otros modelos de IA que se centran en el análisis o la clasificación, la IA generativa produce datos que no existían previamente, imitando patrones aprendidos durante su entrenamiento. Entre sus aplicaciones se incluyen los campos del diseño, entretenimiento, educación, negocios y medicina.

Con relación a la medicina, la IA generativa ayuda, por ejemplo, a la generación de imágenes médicas sintéticas para entrenar sistemas de diagnóstico o en la generación de videoclips personalizados para la rehabilitación de pacientes con afasia, objetivo de este TFG [30].

Para la elaboración de este proyecto, es necesario el estudio del estado del arte de las IA generativas de vídeos, de audios y de movimientos labiales para lograr la sincronización de las anteriores (ver Figura 8).

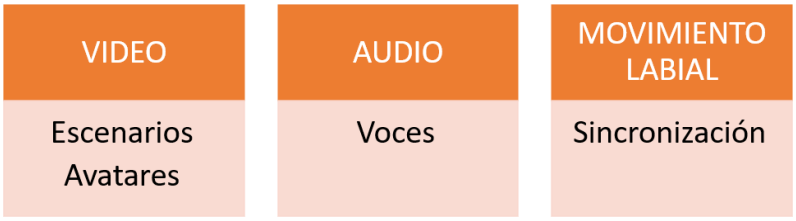

Figura 8. Representación esquemática de cada IA y su función de interés para este TFG.

2.2.1. GENERACIÓN DE VÍDEOS

Es posible generar vídeos a partir de un texto o de una imagen.

- Texto a vídeo: Permite generar vídeos a partir de descripciones textuales, conocidas como prompts, instrucciones o conjunto de palabras proporcionadas para que la IA realice una tarea específica, en este caso, generar vídeo. Los modelos de IA interpretan el texto proporcionado y crean secuencias de vídeo que representan el contenido descrito. Este proceso implica el uso de algoritmos avanzados de procesamiento de lenguaje natural y generación de imágenes en movimiento.

- Imagen a vídeo: Permite generar vídeos a partir de una imagen estática. La IA analiza la imagen y produce una secuencia de vídeo que expande el contenido visual, añadiendo movimiento y dinamismo al elemento original. Este enfoque es útil para animar fotografías o ilustraciones [32].

Actualmente, existe una gran cantidad de herramientas de IA para la generación de vídeos con diversas características y aplicaciones. Se deben evaluar teniendo en cuenta parámetros como su accesibilidad, facilidad de uso, precio, realismo, calidad del contenido generado, personalización e integración en futuros proyectos.

RUNWAY

Runway [33] es una plataforma que ofrece herramientas de IA para la creación de contenido multimedia, incluyendo la generación de vídeos a partir de texto o imágenes. La versión más reciente disponible es Runway Gen-2, que permite a los usuarios crear vídeos de hasta 4 segundos. Sin embargo, en junio de 2024, Runway presentó Gen-3 Alpha, una nueva generación de su modelo de IA capaz de generar clips de vídeo de 10 segundos.

Ofrece una prueba gratuita con limitaciones en la resolución y duración de los vídeos, así como en el almacenamiento disponible (5 GB y 3 vídeos exportables). Los planes de pago comienzan en $12 al mes y permiten acceder a funcionalidades avanzadas como exportaciones de mayor calidad y duración, almacenamiento ampliado de 100 GB, eliminación de marcas de agua, y proyectos de edición ilimitados en el plan Standard. El plan Pro incluye, por $28 al mes, almacenamiento hasta 500 GB y la posibilidad de crear voces personalizadas para Lip Sync y Text-to-Speech [33].

Aunque esta opción podría ser útil en el proyecto, el uso de la interfaz de programación de aplicaciones (API, por sus siglas en inglés) de Runway para integrar el modelo de generación de vídeo en aplicaciones externas requiere un plan de pago y la compra de créditos, que tienen un costo de $0.01 por crédito. Los vídeos generados mediante la API se tarifican según su duración y modelo utilizado. Por ejemplo, un vídeo de 5 segundos generado con Gen-3 Alpha Turbo (resolución 720p) cuesta 25 créditos ($0.25), mientras que un vídeo de 10 segundos cuesta 50 créditos ($0.50).

Para gestionar el acceso a la API, Runway utiliza un sistema de niveles o tiers que limita la cantidad de generaciones diarias y la concurrencia. Por ejemplo, el Tier 1 permite hasta 50 generaciones al día con una concurrencia de 1, mientras que niveles superiores, como el Tier 5, permiten hasta 5,000 generaciones diarias y una concurrencia de hasta 8. El avance entre niveles depende del historial de compras y uso de créditos [34].

STABLE VÍDEO DIFFUSION

Stable Vídeo Diffusion [35] es un modelo de IA desarrollado por Stability AI, diseñado para la generación de imágenes a partir de descripciones textuales. Recientemente, Stability AI ha expandido las capacidades de Stable Diffusion al ámbito de la generación de vídeos mediante el lanzamiento de Stable Vídeo Diffusion, un modelo de código abierto que permite crear vídeos a partir de imágenes.

Stability AI ofrece una API para acceder a sus modelos, incluyendo Stable Vídeo Diffusion, facilitando su integración en aplicaciones y proyectos. La API está disponible tanto para usuarios gratuitos como de pago, aunque los planes no gratuitos ofrecen beneficios adicionales. El acceso gratuito permite realizar un número limitado de solicitudes, con restricciones en el rendimiento y en la resolución de los vídeos generados. Por otro lado, los planes de pago, basados en un sistema de créditos, permiten a los usuarios ajustar la calidad y cantidad de vídeos generados. Cada crédito tiene un costo aproximado de $0.01, y un vídeo corto de 4 segundos puede requerir entre 30 y 50 créditos dependiendo del modelo y la resolución. Este modelo genera vídeos cortos de entre 2 y 4 segundos, con resoluciones de hasta 576x1024 píxeles, a partir de una única imagen de entrada. Sin embargo, la gran potencia computacional necesaria incluso en vídeos de corta duración y en planes no gratuitos, sumado al hecho de que sea un modelo image-to-vídeo, donde se limita a transformar una imagen dada en una secuencia animada, han sido factores limitantes en la participación de esta herramienta en el proyecto [35] [36] [37].

HEYGEN

HeyGen [38] es una plataforma de IA para la creación de vídeos mediante la generación de avatares, lo que la hace especialmente aplicable en marketing o comunicación empresarial. Permite convertir texto en vídeos con avatares que hablan de manera natural, ofreciendo más de 300 voces en más de 40 idiomas. Además, permite la traducción de vídeos y la personalización de avatares mediante la carga de fotografías.

Ofrece una prueba gratuita con ciertas limitaciones que permite crear 3 vídeos al mes de hasta 3 minutos con una calidad de hasta 720 píxeles. Los planes de pago comienzan en $24 al mes, si es facturado anualmente, e incluyen características como eliminación de marcas de agua, acceso a más avatares y voces, mayor duración de los vídeos y exportaciones en mejor calidad.

Sin embargo, para automatizar el proceso y tener acceso a la API, se debe contratar otro plan adicional basado en créditos y donde el más barato es el API Pro-Plan. Pagando $99 al mes incluye 100 créditos, lo que equivale a 1,000 minutos de transmisión de avatares.

En conclusión, esta herramienta presenta algunas limitaciones importantes por las que no ha sido incluido en el proyecto. Los vídeos generados no incluyen escenarios dinámicos y el movimiento de los avatares se restringe a gesticulaciones básicas frente a la cámara, sin interacción con el entorno.

Además, la personalización avanzada, como la creación de avatares únicos, está limitada a los planes más costosos, como el plan profesional. Por último, el uso de la API para integraciones externas requiere un presupuesto adicional [38].

SYNTHESIA

Synthesia [39] es una plataforma líder en la creación de vídeos profesionales generados por IA, que permite convertir texto en vídeos con avatares realistas doblados en más de 140 idiomas. Es tan fácil como crear una presentación de diapositivas por lo que es especialmente útil para aplicaciones en marketing y comunicación empresarial.

Synthesia ofrece un plan gratuito que permite generar hasta 36 minutos de vídeo al año, con acceso a 9 avatares de IA y 2 avatares personales de stock. Además, cuenta con más de 140 idiomas y voces.

En cuanto a los planes de pago, por $18 al mes, si es facturado anualmente, existe el Plan Starter, que ofrece 120 minutos de vídeo al año, acceso a más de 125 avatares de IA y la posibilidad de crear hasta 3 avatares personales. El Plan Creator y el Plan Enterprise, a partir de $64 al mes, mejoran aún más sus funcionalidades y permiten el acceso a la API de Synthesia, no disponible para planes inferiores [39].

Debido a las restricciones de la API y las limitaciones en la personalización de avatares, así como la falta de escenarios dinámicos y la interacción limitada de los avatares, esta herramienta no se ha incluido en el proyecto [40].

HAILUO

MiniMax [42] es una empresa líder en el desarrollo de modelos lingüísticos a gran escala. Hailuo AI es el modelo diseñado para la generación de vídeos de hasta 6 segundos a partir de texto o imágenes, para el cual MiniMax proporciona la infraestructura técnica. Admite la generación de vídeos de alta definición con una resolución de 720 píxeles y 25 fotogramas por segundo.

Hailuo AI utiliza un sistema de créditos flexible que se adapta a diferentes niveles de uso. Los usuarios pueden optar por suscripciones mensuales o por el modelo Pay-as-you-go, que les permite adquirir créditos según sus necesidades sin comprometerse a una suscripción recurrente. En este modelo, los créditos se pueden comprar en incrementos de $5 y tienen una validez de 2 años desde la fecha de compra, brindando una opción adicional para quienes tienen necesidades puntuales de generación de vídeos y no desean suscribirse a un plan mensual.

En cuanto a los planes mensuales, ofrece un plan gratuito con créditos limitados para experimentar con las funcionalidades básicas, pero sin acceso a la API. El plan Standard, por $9.99 al mes en promoción, incluye 1000 créditos iniciales más 100 créditos diarios que expiran si no se utilizan, alcanzando un total de 4000 créditos al mes. Esto equivaldría a unos 130 clips de vídeo al mes aproximadamente. El plan Unlimited, por $94.99 al mes, permite generar vídeos de forma ilimitada con acceso prioritario para acelerar la producción [41].

El uso de la API de Hailuo AI, diseñada para integrar sus capacidades de generación de vídeo en aplicaciones externas, está restringido a usuarios con suscripciones avanzadas, es decir, no está disponible para los planes gratuitos. Cada vídeo generado mediante la API tiene un costo de $0.43 por clip de 6 segundos, lo que incluye características avanzadas como movimientos de cámara cinematográficos [42].

D-ID

D-ID [44] es una plataforma de IA especializada en la creación de vídeos generados a partir de texto e imágenes, permitiendo la animación de rostros y la generación de avatares digitales realistas en múltiples idiomas manteniendo la sincronización labial.

Ofrece un plan de prueba gratuito que dura 14 días, con funcionalidades como la creación de vídeos de hasta 5 minutos, control de expresiones y ajuste de voz, soporte limitado, marca de agua en toda la pantalla y sin acceso a la API. El plan Lite, que consiste en $4.7 al mes, mejora las condiciones, pero sigue sin tener acceso a la API. Ésta está diseñada para desarrolladores y empresas que necesitan personalizar procesos y automatizar la creación de contenido audiovisual, por lo que solo está disponible en los planes Launch, Scale y Enterprise, a partir de $35 al mes. Entre las principales capacidades de la API, gestionada a base de créditos según el plan contratado, se encuentran la generación de vídeos personalizados a partir de texto o imágenes, ajustes avanzados como expresiones faciales y sincronización labial, y la creación de presentadores personalizados con avatares y voces clonadas. Además, permite la traducción y subtitulación automática en múltiples idiomas, lo que podría ser interesante para el proyecto. Sin embargo, debido a la restricción de acceso a la API y a que, aunque la plataforma permite generar vídeos de alta calidad, los avatares generados son efectivos únicamente para presentaciones, con interacción limitada al entorno (restringiéndose a expresiones faciales y gesticulaciones básicas), no se ha incluido D-ID en el proyecto [43] [44].

PIKALABS

Pika Labs [45] es una plataforma de IA que permite la generación de vídeos de alta calidad de entre 3 y 10 segundos a partir de texto e imágenes, ofreciendo diversas funcionalidades para creadores de contenido y desarrolladores.

Ofrece una variedad de planes para adaptarse a diferentes necesidades y niveles de uso. El gratuito permite acceso a Pika 1.5 y 150 créditos mensuales para la generación de vídeos sin marca de agua. El plan Standard asciende a $8 al mes, facturado anualmente, y 700 créditos mensuales. El plan Pro son $28 y 2000 créditos y el plan Fancy $76 y 6000 créditos.

La facturación de la API, que admite hasta 20 generaciones por minuto, se basa en un modelo de pago por uso Pay-as-you-go, facturado mensualmente. Sin embargo, el acceso a la API no está incluido en el plan gratuito ya que está diseñado para usuarios avanzados o empresas, y requiere contactar directamente con el equipo de Pika Labs para su activación [45].

MOCHI 1

Mochi 1 [46] es un modelo de IA de código abierto para la generación de vídeos de hasta 6 segundos a partir de texto desarrollado por Genmo AI, empresa especializada en soluciones de IA generativa. Utiliza la arquitectura Asymmetric Diffusion Transformer (AsymmDiT) para producir vídeos fluidos con movimientos realistas, con una resolución de 480p y 30 fotogramas por segundo, destacándose por su capacidad para adherirse con precisión a los prompts proporcionadas por el usuario.

Ofrece planes para quienes buscan generar vídeos sin necesidad de descargar y ejecutar el modelo localmente. A través del sitio web de Genmo AI, los usuarios pueden acceder al Playground, donde se generan hasta 30 vídeos al mes, 4 cada 6 horas de forma gratuita, con la posibilidad de contratar planes de pago que ascienden a los 80 y 180 vídeos al mes por $8 y $24 al mes respetivamente.

Para los usuarios que necesiten automatizar el proceso, Mochi 1 también ofrece acceso a su API bajo un modelo de Pay-as-you-go, donde los costos se basan en las llamadas realizadas a la API, es decir, se calculan según los créditos consumidos por cada generación de vídeo. Por ejemplo, cada vídeo de 6 segundos requiere 10 créditos, y los usuarios pueden adquirir paquetes de créditos según sus necesidades. El plan Starter incluye 20 créditos por $9.90, equivalente a $4.95 por vídeo, mientras que el plan Pro ofrece 200 créditos por $49, reduciendo el costo por vídeo a $2.45. Para quienes necesitan mayores volúmenes, el plan Premium incluye 800 créditos por $99, con un costo de $1.24 por vídeo.

Mochi 1 está disponible bajo la licencia Apache 2.0, lo que permite a los usuarios descargarlo, modificarlo y usarlo sin restricciones comerciales. Otro de sus grandes beneficios es que no es estrictamente necesario el acceso a la API para utilizar sus funcionalidades principales, ya que, al ser de código abierto, los usuarios pueden descargar el modelo desde Hugging Face y ejecutarlo localmente mediante herramientas como Diffusers, eliminando gastos adicionales y ofreciendo un alto grado de autonomía para quienes cuentan con recursos computacionales suficientes.

Aunque Mochi 1 presenta algunas limitaciones, como la resolución máxima de 480 píxeles en los vídeos generados, se espera que futuras actualizaciones incluyan soporte para HD. Además, muchas características que actualmente están en fase beta o en desarrollo podrían convertirse en funcionalidades plenamente implementadas, lo que reforzará aún más el potencial de esta herramienta [46] [47] [48].

VEO 2

Veo 2 [49] es la nueva herramienta de generación de vídeos mediante IA desarrollada por Google DeepMind. Esta plataforma permite crear vídeos de alta calidad a partir de descripciones textuales, ofreciendo una resolución de hasta 4K, superando a competidores como Sora de OpenAI, que se limita a 1080p, y una comprensión avanzada de la física y el movimiento humano.

Permite la generación de vídeos de hasta 2 minutos, aunque las pruebas actuales de la plataforma de VídeoFX están limitadas a 8 segundos con una resolución de 720 pixeles. Estas limitaciones se deben a que la herramienta aún se encuentra en fase experimental y su acceso está restringido a través de una lista de espera gestionada por Google. Los usuarios interesados en esta herramienta deberán inscribirse en la lista de espera y estar atentos a futuras actualizaciones sobre su disponibilidad general.

Veo 2 promete ser una herramienta revolucionaria en la generación de vídeos, destacándose no solo en la calidad y duración de sus producciones, sino también por su capacidad para simular de manera convincente la física del mundo real. Además, ofrece una amplia variedad de estilos visuales, minimizando artefactos y mejora significativamente la fidelidad de los vídeos generados, incorporando controles de cámara avanzados [49].

2.2.2. GENERACIÓN DE AUDIOS

La generación de audios mediante IA permite, entre otros servicios, simular voces humanas con alta precisión. Estas herramientas se utilizan en diversos ámbitos como el doblaje, asistentes virtuales, narración de vídeos, y en aplicaciones educativas y terapéuticas.

Actualmente, existe una gran cantidad de herramientas de IA para la generación de voces con diversas características y limitaciones. Se deben evaluar teniendo en cuenta parámetros como el idioma, el acento, la pronunciación, la fluidez, la entonación, la velocidad, el tono de voz, el volumen y las pausas.

AZURE SPEECH

Azure Speech Services [50] es una plataforma desarrollada por Microsoft que permite integrar servicios de IA como la conversión de texto a voz (Text-to-Speech, TTS), voz a texto (Speech-to-Text, STT), y traducción de voz en tiempo real, entre otras funcionalidades.

Ofrece voces naturales y personalizables en más de 140 idiomas con un gran número de voces con acento español de España (es-ES), pero también de otras variantes como el español de los países de América del Sur. Permite modular la velocidad, el tono y el estilo o emoción de la voz, lo que permite adaptarlas a diferentes situaciones según el contexto deseado.

El plan gratuito permite procesar hasta 5 horas de audio al mes o hasta 0,5 millones de caracteres para TTS, suficiente para pruebas iniciales y proyectos pequeños. Sin embargo, las voces y estilos disponibles están limitados, y ciertas características avanzadas, como la personalización de voz, no están incluidas. Existe un plan Estándar que ofrece acceso completo a las funcionalidades, incluido el acceso a la API, con precios basados en el uso, es decir, está sujeto al modelo de pago Pay-as-you-go, lo que significa que los costos dependen directamente de la cantidad de procesamiento de voz que se realice. Por ejemplo, la conversión neuronal de texto a voz cuesta $15 por cada millón de caracteres. Además, el plan permite el acceso a voces de alta definición (HD), una mayor variedad de estilos emocionales y herramientas de personalización, como la creación de modelos de voz adaptados a necesidades específicas [50].

ELEVENLABS

ElevenLabs [51] es una herramienta que utiliza IA para ofrecer servicios avanzados de síntesis de voz, como la conversión de texto a voz (Text-to-Speech, TTS), creación de voces a partir de muestras de audios, cambio de voces a modo de doblaje o integración de efectos de sonido.

Es una plataforma muy potente debido a la gran variedad de idiomas a los que ofrece servicio y a la accesibilidad de sus funcionalidades. El plan gratuito permite convertir hasta 10,000 caracteres por mes y brinda acceso a tres voces personalizadas. Además, ofrece acceso a la API sin necesidad de contratar un plan de pago. Sin embargo, ElevenLabs puede limitar el acceso a ciertas voces nativas de idiomas específicos en el plan gratuito, como las voces en español, lo que hace que algunas voces predeterminadas, como las de inglés, intenten leer el texto en español, generando pronunciaciones incorrectas. Este fue el motivo principal por el que no se incluyó en el proyecto, ya que, al querer automatizar la generación de las voces mediante el uso de la API, no era posible la elección de voces en español.

Para hacer frente a estas limitaciones sería necesario contratar el plan Starter por $5 al mes, que ofrece 30,000 caracteres mensuales, acceso a la clonación de voz instantánea y licencia comercial o el plan Creator, a $22 al mes, incluyendo 100,000 caracteres mensuales (~2 horas de audio) y hasta 30 voces personalizadas [51].

HAILUO

La API de Hailuo AI [42], comentada anteriormente en la generación de vídeos, también ofrece opciones de conversión de texto a audio, T2A en la página de MiniMax. Esta opción está igualmente restringida a usuarios con suscripciones avanzadas, es decir, no está disponible para los planes gratuitos. Cada audio generado mediante la API tiene un costo de $30 por millón de caracteres [42].

2.2.3. GENERACIÓN DE MOVIMIENTOS LABIALES

La sincronización labial con IA busca alinear los movimientos faciales, especialmente los labios, de vídeos con audios proporcionados. La principal diferencia con las herramientas de DeepFake es que, mientras estas generan rostros enteros manipulando la apariencia de las personas, las técnicas de sincronización de labios o Lip Syncing se centran en la sincronización labial de un vídeo dado con audios preexistentes.

Diversas plataformas de IA generativas de vídeos como Synthesia o D-ID, mencionadas anteriormente, presentan funcionalidades de DeepFake. Sin embargo, las herramientas de sincronización de labios se adaptan mejor a las necesidades del proyecto, pues es necesario que los personajes parezcan decir palabras o frases, pero sin necesidad de modificar o remplazar sus rostros o identidad.

WAV2LIP

Wav2Lip [52] es un modelo de sincronización labial mediante el uso de IA desarrollado por Rudrabha Mukhopadhyay y su equipo. Se trata de una herramienta de código abierto disponible gratuitamente en GitHub. Utiliza un generador basado en GAN, Generative Adversarial Network, que toma como entrada un vídeo base y un archivo de audio, y produce un vídeo en el que los labios del sujeto coinciden con el audio.

La resolución del vídeo generado con Wav2Lip puede estar limitada por los recursos computacionales disponibles, y en el caso de grabaciones más extensas, podrían aparecer artefactos o inconsistencias visuales. Sin embargo, para este proyecto, donde se requiere una alta precisión y la duración de los vídeos y audios no superará los 6 segundos debido a las limitaciones de las IAs de generación de vídeos, Wav2Lip se presenta como una opción bastante adecuada [52].

3. MÉTODOS E IMPLEMENTACIÓN

3.1. ELECCIÓN DE HERRAMIENTAS

Las soluciones de vídeo basadas en IA siguen presentando dificultades en la creación de escenas de larga duración y alta calidad comparado con las producciones profesionales. Además, muchas de estas herramientas presentan limitaciones añadidas como baja contextualización de escenas, realismo limitado e imposibilidad de generar audio.

Para preservar el alcance emocional, la coherencia y continuidad de los vídeos, y garantizar la inmersión y efectividad del paciente en la terapia de rehabilitación, es necesario la combinación de las tres IA generativas; de vídeo, de audio y de sincronización de ambas.

Teniendo en cuenta las limitaciones mencionadas y a los argumentos posteriormente expuestos, las herramientas de IA seleccionadas para este TFG son Genmo Mochi 1 AI, Azure AI Speech y Wav2Lip (ver Figura 9).

Figura 9. Representación esquemática de las herramientas de IA elegidas para este TFG.

3.1.1. GENMO MOCHI 1

De las herramientas de generación de vídeos estudiadas anteriormente, Genmo Mochi 1 y Stable Vídeo Diffusion son las únicas dos de código abierto, que ofrecen modelos preentrenados descargables y ejecutables sin depender del acceso a su API. Las demás plataformas están diseñadas como servicios cerrados, es decir, si se quiere automatizar el proceso de elaboración de vídeos es necesario poder acceder a su API, y la mayoría son a través de planes de pago.

En cuanto a sus características funcionales, Genmo Mochi 1 presenta una resolución de 480 pixeles, menor resolución comparada con los 576x1024 píxeles que ofrece Stable Vídeo Diffusion. Sin embargo, Mochi 1 es preferible para experimentación y proyectos más ligeros ya que, aunque los requerimientos computacionales son moderados, son igualmente menores a los necesarios para ejecutar Stable Vídeo Diffusion. Además, mediante la realización de pruebas en sus interfaces web, se puede observar cómo los vídeos generados con el playground de Mochi 1 muestran más dinamismo y realismo que otras plataformas donde el movimiento de los avatares se limita a gestos o expresiones en lugar de acciones o movimientos más elaborados. La calidad de movimiento y la adherencia a los prompts proporcionados son dos de las funcionalidades más sobresalientes de este modelo.

- Adherencia a las instrucciones: Demuestra una alineación excepcional con las instrucciones textuales, garantizando que los vídeos generados reflejen con precisión las instrucciones dadas. Esto permite a los usuarios un control detallado sobre los personajes, los ajustes y las acciones. Se evalúa la adherencia a las instrucciones con una métrica automatizada utilizando un modelo de lenguaje de visión como juez siguiendo el protocolo de OpenAI DALL-E 3. Se evalúan los vídeos generados utilizando Gemini-1.5-Pro-002 (ver Figura 10).

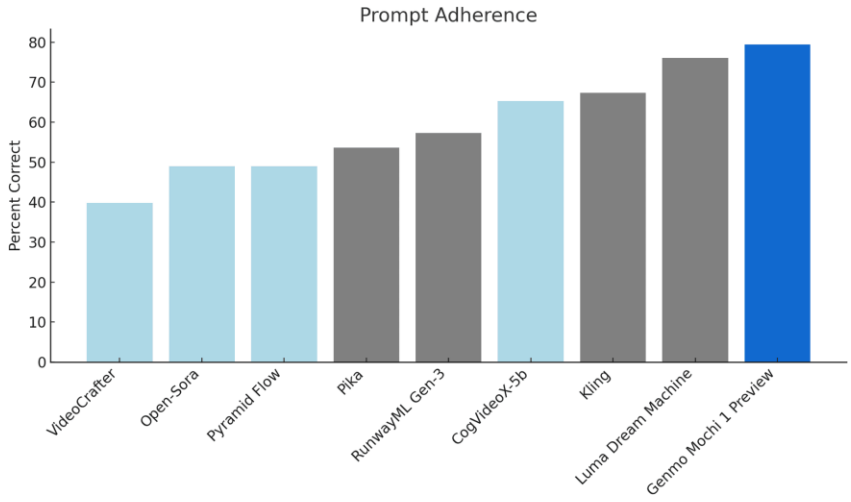

Figura 10. Mide la precisión con la que los vídeos generados siguen las instrucciones textuales proporcionadas, comparándola con otras herramientas de IA generativa de vídeos [46].

- Calidad de movimiento: Mochi 1 genera vídeos fluidos a 30 fotogramas por segundo para duraciones de hasta 6 segundos, con alta coherencia temporal y dinámicas de movimiento realistas. Mochi simula la física, como la dinámica de fluidos y la imitación de pelo y piel, y expresa una acción humana coherente y fluida. En la evaluación de esta función se tuvo en cuenta el movimiento y no en la estética del fotograma (los criterios incluyen el interés del movimiento, la verosimilitud física y la fluidez). Las puntuaciones Elo se calculan siguiendo el protocolo LMSYS Chatbot Arena (ver Figura 11).

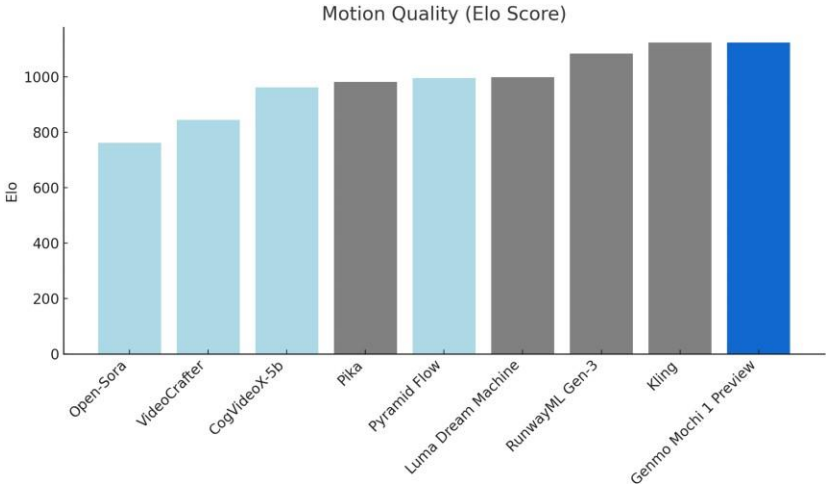

Figura 11. Evalúa tanto la suavidad del movimiento como el realismo espacial, comparándola con otras herramientas de IA generativa de vídeos [46].

En cuanto a la arquitectura del modelo Mochi 1, se conoce como Asymmetric Diffusion Transformer (AsymmDiT) y está basada en un modelo de difusión con más de 10 mil millones de parámetros. Este modelo, entrenado desde cero, es el más grande liberado públicamente para la generación de vídeos y presenta una arquitectura simple y fácilmente modificable.

Mochi 1 incluye un VAE (Vídeo Variational Autoencoder) de código abierto que comprime los vídeos a un tamaño 96 veces menor. Esta compresión se realiza en los ejes espacial (8x8) y temporal (6x) hacia un espacio latente de 12 canales, lo que reduce significativamente el tamaño de los datos procesados.

Procesa prompts junto con los tokens de vídeo comprimidos de manera eficiente, mejorando la capacidad de la red en el razonamiento visual. Optimiza la representación visual al presentar casi 4 veces más parámetros que la corriente textual. Combina texto y tokens visuales mediante self-attention multimodal con capas MLP separadas para cada modalidad.

Además, a diferencia de otros modelos, Mochi 1 emplea un único modelo de lenguaje T5-XXL para codificar los prompts, lo que reduce la complejidad y los costos computacionales que implicaría integrar múltiples modelos de lenguaje y garantiza que los prompts textuales se interpreten de manera uniforme, evitando inconsistencias que podrían surgir al combinar varios modelos.

Atiende a una ventana de contexto de 44,520 tokens con atención tridimensional, es decir, analiza y relaciona una cantidad significativa de datos visuales y textuales en el espacio, lo que resulta en vídeos más coherentes y mejor contextualizados para asegurar la continuidad de las escenas.

Por último, extiende los embeddings posicionales rotatorios (RoPE) a tres dimensiones para manejar interacciones complejas entre los elementos visuales y temporales de un vídeo, ayudando al modelo a entender patrones complejos que involucran cambios visuales y temporales, como movimientos de cámara o acciones rápidas en un vídeo [46].

3.1.2. AZURE SPEECH

Azure AI Services incluye Speech, una herramienta de IA que incorpora el modelo Text-to-Speech (TTS) utilizado en la creación de voces de este TFG.

A pesar de las limitaciones que presenta esta herramienta en el plan gratuito en cuanto a la disponibilidad de ciertas voces y estilos, es la más adecuada y sencilla en términos de accesibilidad e implementación. Ofrece diversas funcionalidades y parámetros modificables con el fin de personalizar y adaptar las voces generadas a los contextos oportunos. Para facilitar esta personalización, se utiliza el archivo de configuración “Voices.json” que permite explorar las características de todas las voces disponibles, incluyendo detalles como región, idioma, estilos y parámetros adicionales. Por ejemplo, para las voces del proyecto, la región donde se sitúa la voz es “westeurope”, el idioma es “es-ES” y la persona y el estilo dependen tanto de las características del personaje, como del género, y la emoción que se quiera transmitir. Si bien una de las restricciones del plan gratuito de Azure consiste en procesar solamente hasta 5 horas de audio al mes, no afecta directamente a este TFG ya que no es necesaria la creación de audios tan largos, si no de audios que incluyan frases simples. Sin embargo, una de las debilidades que más influye es la limitación de estilos en voces de idiomas distintos al inglés. En el caso del español, solamente la voz masculina de Álvaro tiene acceso a estilos emocionales más allá del neutro, como alegre y triste, que se escriben como “cheerful” o “sad” respectivamente.

En cambio, esta limitación se puede contrarrestar modulando la prosodia, con atributos como “prosody”, para simular la fluidez y entonación emocional de las voces. “Prosody” incluye parámetros como “rate”, “pitch”, “volume”, “emphasis”, “break”, “contour” y “duration”. Mediante su ajuste y combinación es posible simular una entonación más natural y fluida, incluso en voces que no tienen estilos habilitados.

- **Rate**: Controla la rapidez con la que se pronuncia el texto. Se puede especificar mediante porcentajes, por ejemplo, +10% o -10% para un 10% más rápido o lento respectivamente, o mediante palabras clave como x-slow, slow, medium, fast y x-fast.

- **Pitch**: Ajusta el tono de la voz, es decir, que tan aguda o grave suena. Se puede definir en semitonos, por ejemplo, +2st o -2st para subir o bajar 2 semitonos respectivamente, o mediante

palabras clave como x-low, low, medium, high y x-high. Al disminuir el tono, la voz sonará más profunda, lo que puede ser adecuado para transmitir seriedad o autoridad.

- **Volume**: Determina el nivel de volumen de la voz sintetizada. Se puede especificar en decibelios relativos, por ejemplo, +3dB para aumentar 3 decibelios, o mediante palabras clave como silent, x-soft, soft, medium, loud y x-loud.

- **Emphasis**: Indica al sintetizador que debe poner mayor o menos énfasis en una palabra o frase específica. Se puede definir mediante los valores strong, moderate y reduced para un énfasis fuerte, moderado o reducido respectivamente. Aplicar un énfasis fuerte en una palabra hará que dicha palabra destaque más en la locución, captando la atención del oyente.

- **Break**: Introduce una pausa en la locución, permitiendo controlar el ritmo y la naturalidad del discurso. Se puede definir en milisegundos, por ejemplo, 500 ms para medio segundo, o mediante palabras clave como none, x-weak, weak, medium, strong y x-strong.

- **Contour**: Permite definir cambios específicos en el tono a lo largo de una frase, creando variaciones más complejas en la entonación. Se especifica mediante una serie de puntos que indican el porcentaje del tiempo y el cambio de tono correspondiente, por ejemplo, (50%, +10%) aumenta el tono en un 10% a la mitad del tiempo.

- **Duration**: Especifica la duración total de la pronunciación de una palabra o frase, permitiendo alargar o acortar su enunciación. Se define en segundos o milisegundos. Establecer una duración más larga en una palabra hará que se pronuncie de manera más prolongada, útil para énfasis dramáticos.

Estos parámetros se implementan utilizando Speech Synthesis Markup Language (SSML), un lenguaje basado en XML que se utiliza para personalizar los resultados de la conversión de texto a voz.

3.1.3. WAV2LIP

Wav2Lip [53] es el modelo abierto de sincronización labial que mejor se adapta a las necesidades y recursos de este proyecto. Un estudio comparó sus funcionalidades con otros modelos que enfrentaban problemas como incapacidad de sincronizar con precisión vídeos dinámicos o no estructurados (en “the wild”). Wav2Lip destacó por superar significativamente a métodos previos como LipGAN y Speech2Vid. Su arquitectura consta de un codificador de identidad, un codificador de audio y un decodificador facial, donde se genera cada fotograma individualmente y utiliza un discriminador experto en LipSync para mejorar tanto la calidad visual como la sincronización labial. Este discriminador utiliza un modelo SyncNet que garantiza una precisión del 91% en la detección de errores de LipSync, frente al 56% de modelos previos como LipGan. Además, Wav2Lip introdujo nuevas métricas, como el Lip Sync Error (LSE) y el Confidence Score (LSE-C) para evaluar la precisión y consistencia de la sincronización de manera objetiva, donde obtuvo resultados significativamente mejores incluso en vídeos no estructurados. De manera subjetiva, los vídeos generados por Wav2Lip fueron preferidos en más del 90% de los casos en evaluaciones humanas [53].

3.2. REQUISITOS TÉCNICOS

Los requerimientos técnicos y los recursos computacionales demandantes son una de las grandes limitaciones a las que se enfrenta un usuario al utilizar la inteligencia artificial generativa. La generación de vídeos, audios y sincronización labial, como en el caso de este TFG, requiere modelos avanzados entrenados con grandes cantidades de datos, lo que resulta en altas demandas de hardware y software a la hora de ejecutarlos e implementarlos en nuestra máquina local. Estas demandas pueden ser inasequibles para individuos o proyectos pequeños sin muchos recursos computacionales debido a la gran exigencia de tarjetas gráficas de última generación con capacidades de procesamiento masivo y memoria de vídeo significativa. Se plantea entonces una barrera importante en términos de accesibilidad que limita las posibilidades de experimentar y desarrollar proyectos basados en IA generativa.

Los requerimientos de hardware y software son cruciales para la implementación y automatización del sistema de generación de escenas.

3.2.1. HARDWARE

En términos de hardware, es necesario contar con un entorno que ofrezca acceso a una Unidad de Procesamiento Gráfico (GPU) suficiente y adecuada para manejar las demandas computacionales de estos modelos, especialmente para la generación de vídeos y la sincronización labial, que son procesos intensivos. Estos procesos consumen mucha potencia de cálculo de la GPU ya que implican realizar millones de cálculos en paralelo. Los modelos de IA generativa deben iterar muchas veces sobre los datos para producir resultados precisos y nuevos, incrementando el tiempo necesario para completar cada tarea, especialmente si el hardware no es lo suficientemente potente.

Por ejemplo, para la generación de escenarios, el modelo Mochi 1 procesa una gran cantidad de datos visuales y temporales. Además, al tener 10 mil millones de parámetros, carga redes neuronales masivas y maneja grandes volúmenes de datos de entrada y salida, lo que satura la memoria disponible en el sistema. Estas funcionalidades requieren una GPU con alta capacidad de procesamiento y suficiente memoria RAM de la GPU o Vídeo RAM (VRAM). En este contexto, el repositorio oficial del modelo soporta tanto la operación multi-GPU (dividiendo el modelo a través de múltiples tarjetas gráficas) como la operación single-GPU aunque esta última requiere aproximadamente 60GB de VRAM para ejecutar la versión de mayor calidad. Para proyectos con recursos limitados, Hugging Face ofrece variantes del modelo que requieren menos carga computacional.

La librería Diffusers, desarrollada por Hugging Face, permite trabajar con modelos de difusión de manera estructurada, flexible y optimizada, dividiendo el proceso de generación en pasos de difusión que pueden ejecutarse en paralelo. Cargar el modelo preentrenado de Mochi 1 con Diffusers, permite configurar las capacidades computacionales (GPU o Unidad Central de Procesamiento (CPU)) según los recursos disponibles descargando versiones específicas del modelo. Esto incluye dos configuraciones principales:

- Modelo de alta resolución: Requiere 42GB de VRAM, pero garantiza la mayor calidad de vídeo.
- Modelo de baja resolución: Utiliza la variante bfloat16 del modelo y requiere 22GB de VRAM para ejecutarse. Como resultado, se obtiene una ligera caída en la calidad del vídeo generado.

En este TFG, se ha optado por cargar el modelo preentrenado de baja resolución para hacer frente a las grandes limitaciones computacionales, sin comprometer la compatibilidad con actualizaciones realizadas por los desarrolladores del modelo y su integración con otras librerías necesarias para llevar a cabo este proyecto.

En tareas como la sincronización labial, el modelo debe analizar cada fotograma de un vídeo, mapearlo con el audio correspondiente y ajustar los movimientos labiales en tiempo real. Esto requiere cálculos simultáneos en múltiples dimensiones (espacial, temporal y acústica), lo cual aumenta la carga computacional considerablemente. Para una experiencia óptima es recomendable usar una GPU con ≥ 8 GB VRAM, 16 GB RAM, quedando cubiertas estas necesidades computacionales con los requerimientos técnicos que se van a utilizar para poder cargar el modelo de generación de vídeos de Mochi 1.

3.2.2. SOFTWARE

A nivel de software, la ejecución e integración de los modelos utilizados en este proyecto requieren un entorno especializado que permita optimizar el rendimiento y gestionar eficientemente los recursos disponibles. Para la generación de vídeos, el modelo Mochi 1 se ejecuta mediante la librería Diffusers, comentada anteriormente. Además, el modelo requiere el uso de PyTorch, un framework de aprendizaje profundo que habilita la ejecución eficiente del modelo en GPU con soporte CUDA. En cuanto a la generación de voces y sincronización labial, FFmpeg es necesaria como herramienta adicional para la manipulación y postprocesamiento de archivos de audio. Ajusta la frecuencia de muestreo, convierte el formato del audio a WAV, compatible con la sincronización labial, y finalmente, une los cuatro vídeos que conforman la escena completa. En el caso de Azure, se requiere la instalación de librerías como requests y pydub para gestionar las llamadas a la API y convertir los formatos de audio generados. La autenticación se realiza mediante un token (API Key) generado a partir de una cuenta de estudiante, que permite acceder a las funcionalidades.

Por último, para garantizar la interoperabilidad entre las diferentes etapas del pipeline generado para el sistema de generación de vídeos, se ha utilizado Python como lenguaje principal de programación.

3.2.3. GOOGLE COLAB PRO

Las necesidades computacionales para automatizar este proyecto se resumen en las siguientes.

Hardware:

- **GPU**: Para un rendimiento óptimo se necesitan GPU de gama alta como NVIDIA A100 o V100, capaces de manejar cargas intensivas de procesamiento.
- **RAM**: Un mínimo de 16 GB de memoria RAM para manejar eficientemente los datos generados y procesados durante la ejecución del modelo.
- **Almacenamiento**: Al menos 10 GB de espacio libre, considerando el tamaño del modelo preentrenado, los datos de entrada y salida, y los archivos temporales generados durante la ejecución.

Software:

- **Frameworks**: Compatibilidad con PyTorch y CUDA para aprovechar la aceleración por GPU.
- **Librerías**: Diffusers es necesaria para inicializar y configurar el pipeline del modelo de baja resolución.

Durante el TFG, investigamos la posibilidad de adquirir GPU de 40GB o 60GB, pero debido a su elevado coste, se optó por crear una cuenta en Google Colab Pro para poder abordar estas necesidades, al tener acceso a GPU de alto rendimiento sin la necesidad de adquirir hardware dedicado.

Google Colab Pro es una solución basada en la nube que permite a los usuarios acceder a recursos de GPU y Unidad de Procesamiento Temporal (TPU) para ejecutar proyectos intensivos en computación. Por un coste de 11,19 € al mes, Google Colab Pro ofrece 100 unidades de procesamiento al mes, acceso a GPU más rápidas y potentes como Tesla P100 o A100, ideales para ejecutar y entrenar modelos de IA, y acceso a máquinas con mayor capacidad de memoria, hasta 40 GB de VRAM.

- **GPU más rápidas**: Los usuarios que hayan comprado uno de los planes de pago de Colab tendrán acceso a GPU premium. Se puede cambiar los ajustes de tu cuaderno para usar una GPU superior cambiando el entorno de ejecución y habilitando el acelerador Premium. En función de la disponibilidad, si se selecciona una GPU premium, podrás acceder a una GPU Nvidia L4 o A100. La versión sin coste económico de Colab ofrece acceso a GPUs T4 de Nvidia sujeto a restricciones de cuota y de disponibilidad.

- **Más memoria**: Los usuarios que hayan comprado uno de los planes de pago de Colab tendrán acceso a máquinas virtuales (VM) con alta capacidad de memoria cuando estén disponibles. Se puede habilitar un entorno de ejecución de alta capacidad de RAM cambiado el tipo de entorno de ejecución y eligiendo Alta capacidad de RAM en el menú desplegable Características del entorno de ejecución.

- **Tiempos de ejecución más largos**: Todos los entornos de ejecución de Colab se restablecen tras un periodo concreto, que es más breve si el entorno no está ejecutando ningún código. Los usuarios de Colab Pro y Pro+ tienen acceso a tiempos de ejecución más largos que los usuarios de la versión sin coste económico de Colab.

- **Ejecución en segundo plano**: Los usuarios de Colab Pro+ tienen acceso a la ejecución en segundo plano: al usarla, los cuadernos se seguirán ejecutando, incluso después de que hayas cerrado una pestaña del navegador. Esta función siempre se habilita en los entornos de ejecución Pro+ mientras que tengas unidades de computación disponibles.

El sistema de Colab utiliza unidades computacionales o de procesamiento para distribuir el acceso a los recursos. Determinan la prioridad de acceso a los recursos según la suscripción, teniendo acceso prioritario Colab Pro+ sobre Colab Pro, y éste sobre los usuarios gratuitos. Es importante gestionar las unidades computacionales eficazmente ya que un uso continuado de sesiones largas o tareas muy intensivas puede agotarlas, afectando al rendimiento. Google Colab Pro+ ofrece, por 51,12 € al mes, 500 unidades de procesamiento y acceso prioritario a GPU más potentes. Sin embargo, en este TFG, el plan Pro ha sido suficiente para cargar el modelo de baja resolución de Mochi 1 y hacer el pipeline necesario para la generación completa de las escenas, aunque con menor calidad. Además, Google Colab permite dividir el pipeline del proyecto en diferentes cuadernos o notebooks para ejecutar cada etapa de las IA integradas (generación de vídeos, audios y sincronización labial) de manera eficiente y en un entorno diferente. Esto no solo optimiza el uso de los recursos disponibles, sino que también facilita el manejo modular del sistema [54].

3.3. ARQUITECTURA DEL SISTEMA

Para poder abordar la arquitectura final del sistema de generación de escenas mediante IA, es necesario describir los archivos principales mochi.py y azure.py, posteriormente importados en el entorno de ejecución de Google Colab Pro como entradas para la sincronización labial. Cada archivo cumple una función específica dentro del pipeline, facilitando la automatización de tareas como la creación de escenarios y la generación de voces.

3.3.1. MOCHI.PY

El archivo mochi.py contiene todas las funciones necesarias para la creación de escenarios y personajes utilizando el modelo de IA generativa Mochi 1. En este TFG, se carga la versión de baja resolución del modelo, implementada con la librería Diffusers, lo que permite generar vídeos a partir de prompts configurados dinámicamente. Estos prompts se adaptan en función de los atributos de entrada definidos manualmente por el usuario según las características deseadas para cada escena. El resultado final obtenido al ejecutar el archivo mochi.py consistirá en 4 vídeos diferentes e independientes, cada uno representando una parte del escenario final: el contexto inicial, el contexto final, la interacción principal y la interacción secundaria.

- **Contexto inicial**: Este vídeo muestra la introducción al escenario donde se desarrolla la escena, incluyendo la entrada del personaje principal al entorno general. Se define la atmósfera inicial para hacernos una idea inicial del marco visual que contextualiza la interacción posterior.

- **Interacción principal**: En este vídeo se desarrolla el diálogo entre el personaje principal, el cual sería el interpretado por el paciente en las sesiones de rehabilitación, y un personaje secundario que depende del contexto. Por ejemplo, si se está practicando en el entorno de una consulta médica, el personaje secundario que interactúa con el paciente será un médico, y en el caso de una cafetería o un restaurante, el personaje secundario será un camarero.

- **Interacción secundaria**: Este vídeo se enfoca en la respuesta del personaje secundario a la interacción anterior por parte del personaje principal.

- **Contexto final**: En este vídeo se representa el desenlace o cierre de la escena, incluyendo en algunos casos, la salida del personaje u otra escena a modo de conclusión.

Con el objetivo de mejorar la flexibilidad y la coherencia visual en los resultados generados, se han implementado dos funciones independientes para generar los contextos y las interacciones por separado. Esto se hizo por varios motivos. Por un lado, si los contextos generados no cumplen con las expectativas pueden repetirse sin necesidad de ejecutar otra vez las interacciones. Esto permite ajustar cada vídeo de manera independiente, ahorrando tiempo y recursos, y mejorando la calidad del resultado final. Por otro lado, realizar intencionadamente las agrupaciones ‘contexto inicial-contexto final’ e ‘interacción principal-interacción secundaria’, garantiza al máximo la coherencia y cohesión visual en los 4 vídeos generados, considerando que la IA puede interpretar los prompts de manera diferente. Al generar los contextos juntos, se puede prestar mayor atención a los detalles del entorno y la atmósfera general, garantizando una continuidad visual del lugar. De este modo, se evita proporcionar demasiados detalles sobre los personajes, ya que en estas tomas la IA podría confundirse y prestar excesiva atención a características no relevantes de los mismos, en lugar de centrarse en los elementos del lugar que son clave en el contexto. Por su parte, en las interacciones, donde los personajes son los protagonistas y se aprecian más de cerca, la IA puede enfocarse más en la apariencia física y asegurarse de que se parezcan y se mantenga una consistencia entre las tomas, sin estar tan pendiente de los detalles del lugar.

En la Figura 12 queda definido gráficamente el flujo de trabajo que se sigue en el archivo de Python, mochi.py.

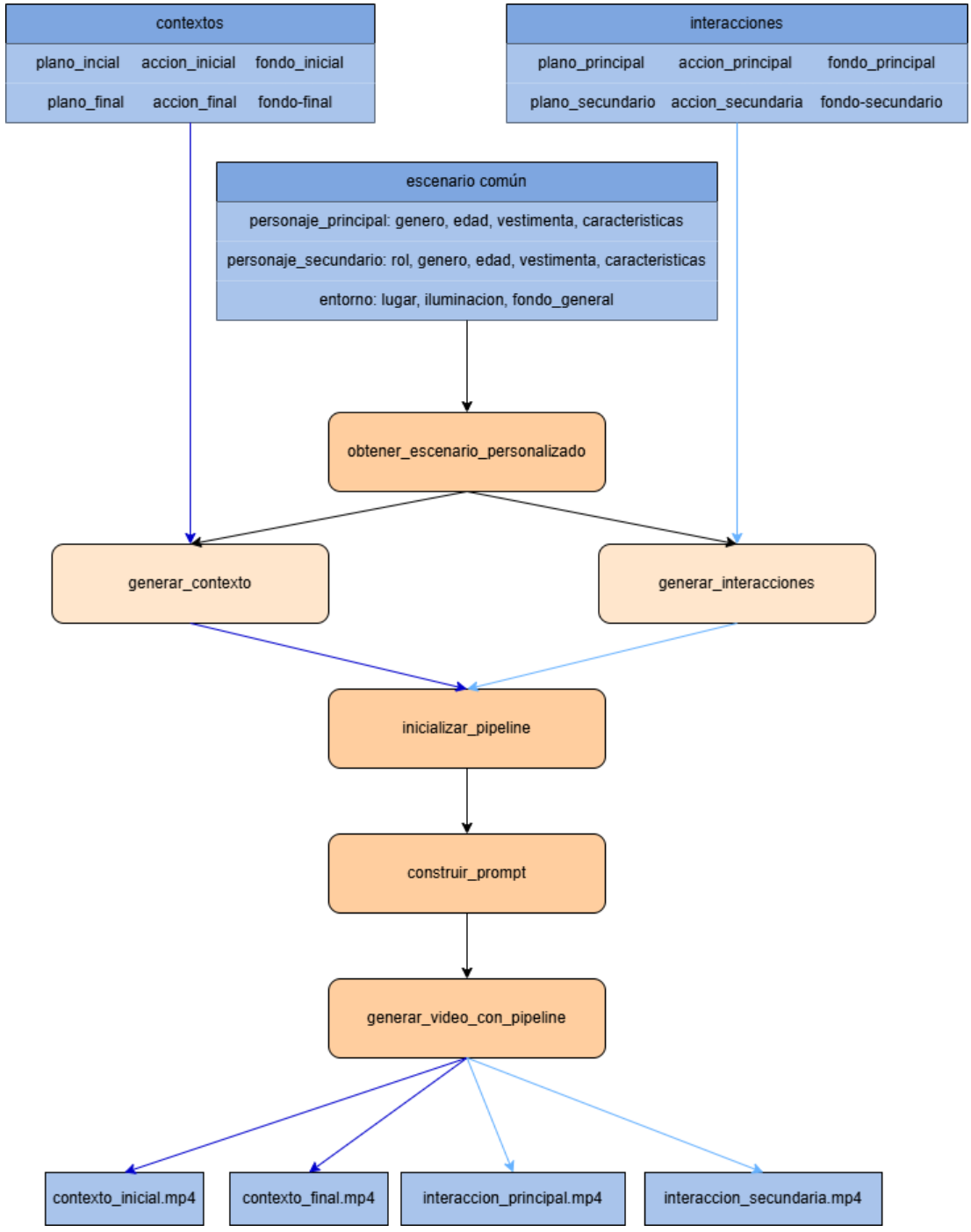

Figura 12. Diagrama del flujo de trabajo seguido para la generación de vídeos en el archivo mochi.py. En azul, se pueden distinguir los atributos de entrada modificables manualmente para cada escena y los vídeos resultantes contexto_inicial, contexto_final, interacción_principal e interacción_secundaria. En naranja, se observan las funciones del archivo y su concatenación. La flecha de color azul oscuro sigue el flujo para la generación de las escenas contextuales y la flecha de color azul claro sigue el flujo para la generación de las escenas donde interactúan directamente los personajes principal y secundario.

Para estructurar y optimizar la generación de los vídeos, se ha diseñado un escenario común que sirve como base para todas las escenas generadas, ya sean contextos o interacciones. Este escenario, definido en la función **obtener_escenario_personalizado()**, incorpora los atributos descritos en la Figura 12 para caracterizar al personaje principal, al personaje secundario y al entorno. En el archivo se incluye una pequeña plantilla con posibles ejemplos para varios escenarios de interés, que podría ser ampliado en un futuro.

```
def obtener_escenario_personalizado():
    return {
        "personaje_principal": {
            "genero": "man",
            "edad": "elderly",
            "vestimenta": "dark green sweater and dark trousers",
            "caracteristicas": "white hair and without glasses"
        },
        "personaje_secundario": {
            "rol": "doctor",
            "genero": "man",
            "edad": "middle-aged",
            "vestimenta": "white coat and black trousers",
            "caracteristicas": "short black hair, wearing a black stethoscope"
        },
        "entorno": {
            "lugar": "doctor's consultation room",
            "iluminacion": "bright and clinical lighting",
            "fondo_general": "a counter with medical equipment and charts"
        }
    }
```

A partir de este escenario común, se pasan los datos al generador de contextos e interacciones definidos en las funciones **generar_contexto(escenario, incluir_secundario=False)** y **generar_interacciones(escenario)**, donde se definen parámetros más específicos para cada parte como el plano de la cámara, las acciones de los personajes y el fondo en cada escena, y se ejecutan el resto de llamadas a las funciones del pipeline.

Una vez definidos estos parámetros, y dentro de las funciones anteriores, se inicializa el pipeline llamando a la función **inicializar_pipeline()**. Este proceso carga el modelo de Mochi 1 desde Diffusers, ajustando las capacidades computacionales según el entorno de ejecución disponible, lo que asegura que el modelo esté listo para procesar los prompts. En este proyecto, debido a las limitaciones de hardware, se ha ejecutado el modelo de baja resolución.

Con el pipeline configurado, se construye el prompt mediante la función **construir_prompt(plano, accion, fondo, escenario, incluir_secundario=False, tipo_interaccion=None)**. Esta función combina los parámetros de entrada para generar una descripción textual detallada del escenario y las acciones deseadas, asegurando que el modelo interprete correctamente las características definidas para la escena. Se pasa como parámetro el escenario con los atributos comunes, y el plano, acción y fondo específico de cada vídeo y definidos anteriormente. El parámetro incluir_secundario es un booleano que permite decidir si el personaje secundario será incluido en la descripción del prompt. Este ajuste es particularmente útil en los vídeos de contexto, donde el enfoque está en las características del entorno y las acciones generales del personaje principal, como su entrada o salida del escenario. Al omitir la

descripción física del personaje secundario (dejando el parámetro como False), se evita que la IA dedique demasiados recursos a detallar al personaje secundario en escenas donde su relevancia es mínima. Esto ayuda a mantener el foco en el entorno o en las características principales del escenario, garantizando mayor coherencia visual y narrativa. Por defecto, este valor es False, pero puede configurarse como True para que el modelo integre al personaje secundario en la escena, junto con sus características previamente definidas. El parámetro tipo_interaccion permite especificar el tipo de interacción que ocurre entre los personajes, especificando si el foco de la escena estará en el personaje principal mientras habla, o en el personaje secundario cuando responde. Esta configuración influye directamente en el plano de la cámara, asegurando que el personaje relevante se encuentre en el centro de la acción y en enfoque, mientras que el otro personaje permanece ligeramente desenfocado en el primer plano (foreground). Este efecto no solo ayuda a dirigir la atención del espectador al personaje activo en la conversación, sino que también aporta un mayor realismo y profundidad a la escena generada. Al combinar todos estos parámetros, la función asegura un alto nivel de personalización y precisión en la generación de vídeos, adaptando el resultado a las necesidades específicas de cada escena y manteniendo una consistencia visual y narrativa en el flujo de trabajo.

```
prompt_inicial = construir_prompt(plano_inicial, accion_inicial, fondo_inicial, escenario, incluir_secundario)
```

```
prompt_final = construir_prompt(plano_final, accion_final, fondo_final, escenario, incluir_secundario)
```

```
prompt_principal = construir_prompt(plano_principal, accion_principal, fondo_principal, escenario, tipo_interaccion="principal")
```

```
prompt_secundario = construir_prompt(plano_secundario, accion_secundaria, fondo_secundario, escenario,tipo_interaccion="secundaria")
```

Finalmente, el vídeo se crea con la función **generar_vídeo_con_pipeline(pipe, prompt, output_filename)**. Esta función utiliza el pipeline inicializado y el prompt construido para cada escena, prompt_inicial, prompt_final, prompt_principal y prompt_secundario. El vídeo generado para cada escena se guarda automáticamente en un formato compatible para su posterior procesamiento, como la sincronización labial o la integración en el sistema completo. Además, la duración del vídeo se puede ajustar modificando número de fotogramas y los fotogramas por segundo siendo num_frames/fps = duración del vídeo, y con un máximo de 6 segundos.

3.3.2. AZURE.PY

El archivo azure.py contiene todas las funciones necesarias de filtrado y generación de voces para los personajes principal y secundario, utilizando el modelo de IA generativa Azure Speech. Para acceder a las funcionalidades de su API gratuita, es necesario crear una cuenta de estudiantes donde habilitar una clave o token, conocida como API Key. Se trata de un código alfanumérico extenso que habrá que guardar para incluir en el parámetro de api_key en el archivo de Python azure.py. Además, es posible definir el idioma y la región a la que pertenezcan las voces que se desean generar. En este proyecto, el resultado final consistirá en dos archivos de audio con la extensión WAV, output_audio1_16k y output_audio2_16k, que corresponderán a la frase del personaje principal con una voz determinada y a la frase del personaje secundario con una voz distinta, ambas adaptadas en tono, velocidad y parámetros definidos previamente en otros capítulos de este trabajo.

En la Figura 13 queda definido gráficamente el flujo de trabajo que se sigue en el archivo de Python, azure.py. Una vez definidos los atributos comunes para ambas voces, como la API Key, la región y el idioma, se describe el género de las voces principal y secundaria, seleccionando entre ‘male’ o ‘female’, dependiendo de si se necesita una voz masculina o femenina. Además, se definen las frases que se escucharán en los audios, representando los diálogos de los personajes principal y secundario durante las escenas de interacción. Estas frases, descritas en los parámetros de texto_principal y texto_secundario, son modificables manualmente, lo que permite adaptarlas a las palabras que se deseen practicar en las sesiones de rehabilitación. En cuanto al estilo, la API gratuita de Azure Speech solo ofrece esta funcionalidad para la voz española y masculina de Álvaro, con opciones limitadas a los estilos ‘cheerful’ y ‘sad’. Por este motivo, el parámetro de estilo solo se aplica al personaje principal, ya que es el interpretado por el paciente y puede beneficiarse de esta personalización. En el caso de que el personaje principal sea femenino, y dado que no existen voces femeninas en español con estilos disponibles en el plan gratuito, se seleccionará automáticamente la primera voz femenina que cumpla con los demás criterios configurados. Por otro lado, para la voz secundaria, se utiliza el parámetro “excluirvoz”, el cual permite excluir la voz asignada al personaje principal si ambos comparten el mismo género. Esto asegura que las voces de ambos personajes sean diferenciadas, favoreciendo una interacción más natural en las escenas generadas.

La función **obtener_voz_filtrada(api_key, region, genero, idioma, estilo=None, excluirvoz=None)**, permite seleccionar la voz más adecuada para cada personaje, en función de los parámetros definidos previamente, como el género, el estilo y las restricciones establecidas por excluirvoz en el caso de la voz secundaria. Si el personaje principal tiene asignado un estilo, se prioriza encontrar una voz que cumpla con esta característica. En el caso de la voz secundaria, se garantiza que sea distinta de la asignada al personaje principal. Si no se logra encontrar una voz con las características descritas, el modelo devuelve un mensaje para comunicarlo.

Una vez filtradas y seleccionadas las voces a utilizar, se generan los audios mediante la función **generar_audio(api_key, region, voz, texto, estilo=None, output_name="output_audio.wav")**. Se toma como entrada el texto definido previamente para cada personaje junto con la configuración de la voz seleccionada. Se generan los archivos de audio correspondientes en varios formatos, como MP3 y WAV, facilitando la integración de los audios etapas posteriores del pipeline. Además, en esta función se define el campo “data”, que permite ajustar parámetros avanzados como la prosodia, ofreciendo una alternativa para modular las voces sin estilo y adecuarlas a las necesidades clínicas.

Como paso final en el proceso de generación de audios, se realiza una conversión adicional al archivo con extensión WAV a una frecuencia de muestreo de 16 kHz mediante la herramienta FFmpeg, utilizando el comando subprocess.run. Este ajuste es esencial para que los audios sean compatibles con Wav2Lip, que requiere esta configuración específica para realizar una sincronización labial precisa entre el audio y los vídeos generados.

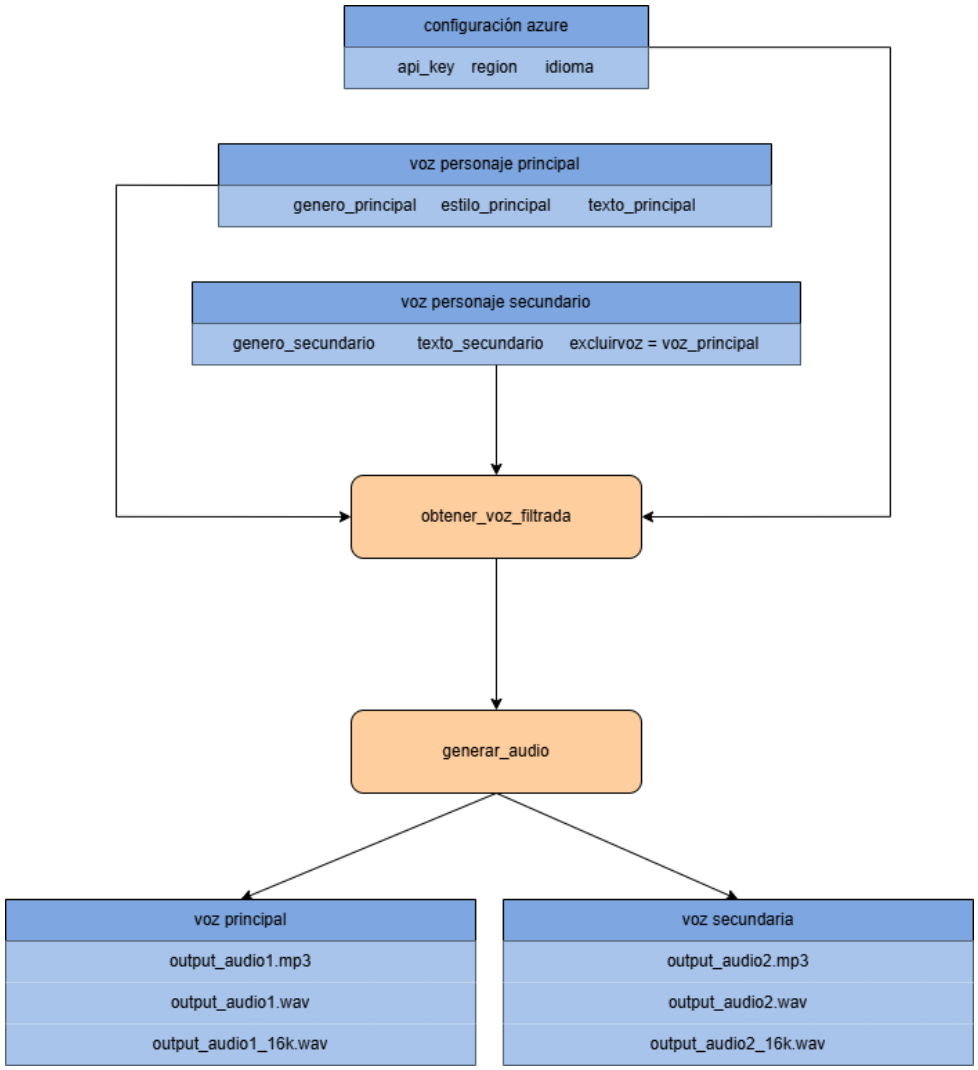

Figura 13. Diagrama del flujo de trabajo seguido para la generación de voces en el archivo azure.py. En azul, se pueden distinguir los atributos de entrada modificables manualmente para cada voz y los audios resultantes output_audio1_16k y output_audio2_16k en el formato necesario para Wav2Lip. En naranja, se observan las funciones del archivo y su concatenación.

3.3.3. IMPLEMENTACIÓN FINAL

En la Figura 14 se puede visualizar la concatenación final de los archivos generados mediante Mochi 1 y Azure Speech, y su implementación en un entorno de ejecución proporcionado por Google Colab Pro.

El pipeline ha sido estructurado en dos notebooks independientes, lo que aporta una arquitectura modular a la automatización del proceso. Esta división se realizó principalmente porque la herramienta de sincronización labial, Wav2Lip, requiere una configuración distinta a las otras herramientas utilizadas. Intentar integrar todo en un único entorno habría resultado en constantes ajustes de configuración, agotando los recursos disponibles e incrementando la complejidad y el tiempo necesario para la ejecución. Además, mantener las tareas separadas en diferentes notebooks mejora la organización general del proyecto y facilita el manejo de posibles errores. Por ejemplo, si se detecta un problema en la sincronización labial, se puede depurar directamente en el notebook correspondiente sin afectar el flujo de generación de vídeos o audios.

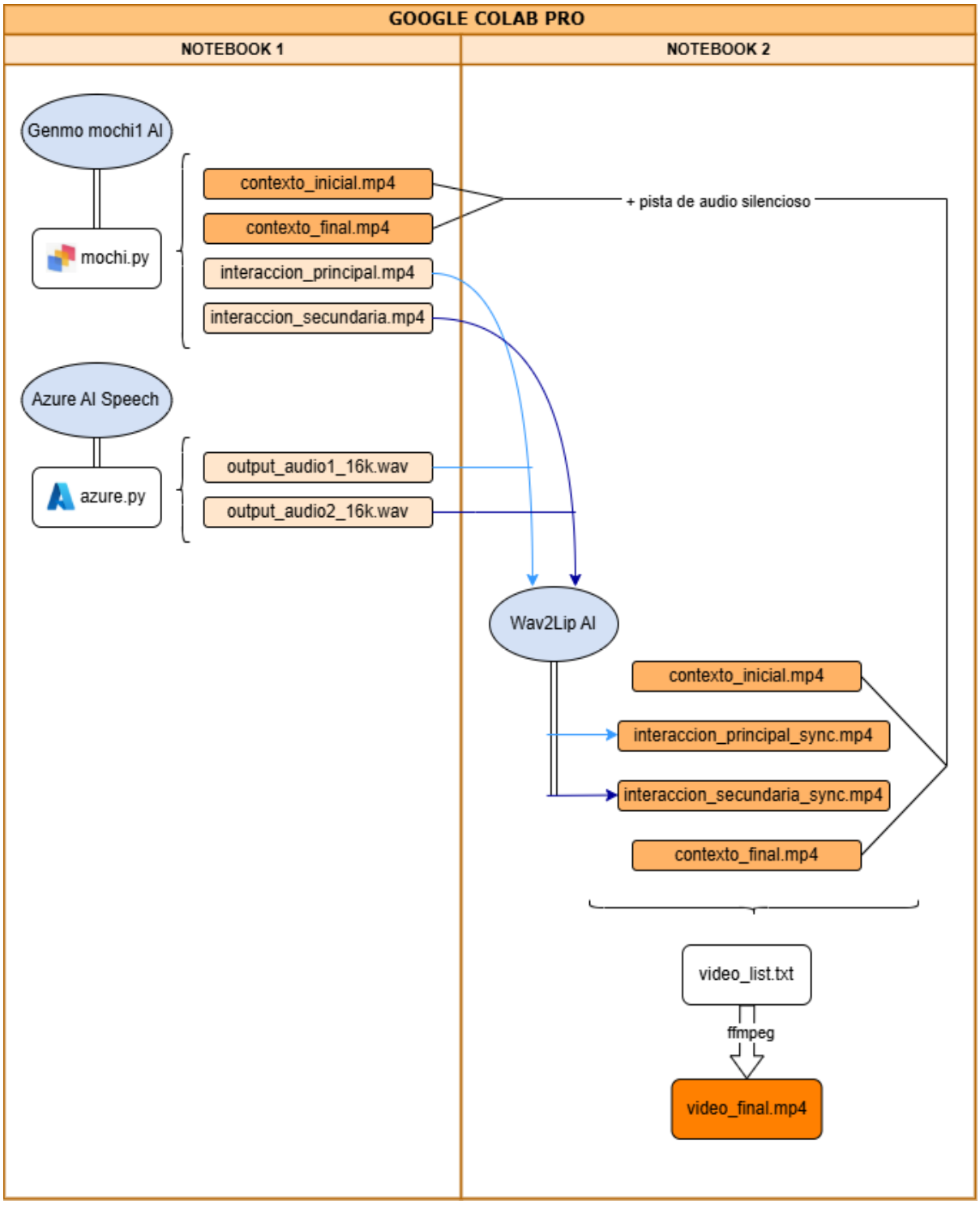

Figura 14. Esquema final del pipeline desarrollado en el sistema de generación de escenas para la rehabilitación de afasia mediante la integración de diferentes IA generativas. La flecha de color azul claro representa el flujo de trabajo seguido para generar el vídeo de la interacción principal, y la flecha de color azul oscuro representa el de la interacción secundaria. El resultado final es un archivo de vídeo denominado vídeo_final.mp4, que consta de 4 escenas donde los personajes hablan dentro de un contexto que se quiere practicar, permitiendo que los pacientes les doblen en las sesiones de rehabilitación.

Como se puede observar en la Figura 14, los escenarios y las voces se van a generar en el primer notebook mientras que, en el segundo, se sincronizarán los labios en las escenas de interacción y se concatenarán los 4 vídeos para obtener el vídeo final.

NOTEBOOK 1

```
!pip install diffusers transformers torch safetensors pydub ffmpeg gdown
!apt-get install -y ffmpeg
```

La primera línea de código instala las librerías necesarias para ejecutar el modelo de generación de vídeos y voces.

- **Diffusers**: Librería de Hugging Face para cargar y ejecutar el modelo de baja resolución de Mochi 1.
- **Transformers**: Librería para trabajar con modelos de lenguaje preentrenados como T5 (utilizado en Mochi 1).
- **Torch**: Framework de DL (PyTorch) que permite ejecutar los modelos en GPU o CPU.
- **Safetensors**: Permite cargar modelos de Hugging Face de forma segura y eficiente en memoria.
- **Pydub**: Herramienta para manipular archivos de audio, como cambiar su formato o ajustarlos para Wav2Lip.
- **Ffmpeg**: Herramienta para procesar audio y vídeo. En el proyecto se utiliza para convertir audios al formato necesario (WAV) y ajustarlos a 16 kHz para la sincronización labial.
- **Gdown**: Herramienta para descargar archivos directamente desde Google Drive, útil para cargar modelos grandes alojados en la nube.

La segunda línea de código instala ffmpeg desde los repositorios de Ubuntu para garantizar que esté disponible en el entorno. Aunque también se instala mediante pip, esta línea asegura compatibilidad con el sistema operativo.

```
!pip install git+https://github.com/huggingface/diffusers.git
```

Instala la versión más reciente del repositorio Diffusers directamente desde GitHub. Esta línea de código es fundamental para cargar y ejecutar el modelo de Mochi 1 posteriormente.

```
%run mochi.py
```

Ejecuta el archivo mochi.py, explicado previamente en la sección 3.3.1. Como resultado, genera los 4 vídeos que representan el contexto inicial, la interacción principal, la interacción secundaria y el contexto final.

```
%run azure.py
```

Ejecuta el archivo azure.py, explicado previamente en la sección 3.3.2. Como resultado, genera los 2 audios correspondientes a las voces de los personajes.

NOTEBOOK 2

Una vez generado el material con el que se va a trabajar, solo falta sincronizar audio y vídeo con la herramienta de Wav2Lip para las interacciones, añadir una pista de audio silenciosa en los contextos y concatenar las 4 escenas.

```
!sudo apt-get update -y
!sudo apt-get install python3.7 python3.7-distutils -y
!sudo update-alternatives --install /usr/bin/python3 python3 /usr/bin/python3.7 1
!sudo update-alternatives --config python3
```

```
!curl https://bootstrap.pypa.io/pip/3.7/get-pip.py -o get-pip.py
```

```
!python3 get-pip.py
```

En estas líneas se instalan las dependencias necesarias. En primer lugar, se actualiza el índice de paquetes de Ubuntu para garantizar que se instalen las versiones más recientes de las dependencias. Se instala Python 3.7 y los módulos para manejar distribuciones. Se añade Python 3.7 como una alternativa para python3, se establece su prioridad y se configura para que use la versión instalada. Se descarga el script de instalación de pip compatible con Python 3.7 y por último, lo ejecuta para instalar pip.

```
!git clone https://huggingface.co/camenduru/Wav2Lip /content/Wav2Lip
```

Clona el repositorio de Wav2Lip desde Huggingface en el directorio /content/Wav2Lip.

```
!pip install numpy==1.19.5 scipy==1.5.4 librosa==0.8.0 numba==0.51.2
!pip install opencv-python==4.5.3.56 opencv-contrib-python==4.5.3.56
!pip install tqdm==4.62.3 ffmpeg-python==0.2.0
!pip install torch==1.10.0+cpu torchvision==0.11.0+cpu --no-cache-dir -f
https://download.pytorch.org/whl/torch_stable.html
```

```
%cd /content/Wav2Lip
```

En estas líneas se instalan las librerías requeridas para el procesamiento de audio y cálculos científicos. Se instala OpenCV y sus contribuciones, necesarias para el manejo de vídeo y la sincronización de labios. Se instalan librerías para mostrar barras de progreso y manejar ffmeg en Python. Por último, se instala PyTorch y torchivision para ejecutar modelos de DL, configurado para CPU.

Cambia el directorio actual a /content/Wav2Lip para ejecutar la sincronización.

```
!python3 inference.py --checkpoint_path checkpoints/wav2lip_gan.pth \
                      --face /content/interaccion_principal.mp4 \
                      --audio /content/output_audio1_16k.wav \
                      --outfile /content/interaccion_principal_sync.mp4
```

```
!python3 inference.py --checkpoint_path checkpoints/wav2lip_gan.pth \
                      --face /content/interaccion_secundaria.mp4 \
                      --audio /content/output_audio2_16k.wav \
                      --outfile /content/interaccion_secundaria_sync.mp4
```

Estas líneas de código ejecutan el script de inferencia de Wav2Lip utilizando el archivo wav2lip_gan.pth que contiene los pesos entrenados del modelo. Especifica el archivo donde se analizarán los rostros y movimientos labiales, es decir, los vídeos generados por Mochi 1, y determina el archivo de audio que se sincronizará con los labios del personaje, es decir, los vídeos generados por Azure Speech. Como resultado, se obtendrán dos vídeos separados para cada sincronización.

El archivo interacción_principal_sync.mp4 para el personaje principal y el archivo interacción_secundaria_sync.mp4 para el personaje secundario.

```
%cd /content
```

```
!ffmpeg -i contexto_inicial.mp4 -f lavfi -t 5.43 -i
anullsrc=channel_layout=mono:sample_rate=16000 -shortest
contexto_inicial_audio.mp4
!ffmpeg -i contexto_final.mp4 -f lavfi -t 5.43 -i
anullsrc=channel_layout=mono:sample_rate=16000 -shortest
contexto_final_audio.mp4
```

Se cambia al directorio raíz del trabajo para poder acceder a los archivos contextuales y añadirles una pista de audio silenciosa, necesaria para poder concatenar las 4 partes. Usa ffmpeg para combinar el vídeo de contexto inicial y contexto final con un audio en silencio, en este caso de 5.43 segundos, pero se puede modificar la duración en función de lo que nos interese.

```
!echo "file 'contexto_inicial_audio.mp4'" > vídeo_list.txt
!echo "file 'interaccion_principal_sync.mp4'" >> vídeo_list.txt
!echo "file 'interaccion_secundaria_sync.mp4'" >> vídeo_list.txt
!echo "file 'contexto_final_audio.mp4'" >> vídeo_list.txt
```

```
!cat vídeo_list.txt
```

```
!ffmpeg -f concat -safe 0 -i vídeo_list.txt -c:v copy -c:a aac -strict
experimental vídeo_final.mp4
```

Crea un archivo vídeo_list.txt donde se agrega la ruta de las 4 partes en orden y se verifica que las rutas sean correctas. Por último, se usa ffmeg para concatenar los vídeos de la lista y generar el archivo final con una duración aproximada de 20 segundos llamado vídeo_final.mp4.

4. RESULTADOS

En este capítulo se exponen los resultados obtenidos tras implementar el sistema de generación de vídeos diseñado en este proyecto. Todas las escenas contienen la marca de agua de la IA de Mochi 1 (genmo). Sin embargo, el nombre del archivo se ha incluido únicamente con carácter informativo, es decir, no aparece en los vídeos generados ni en el vídeo final. Todos los escenarios y las frases seleccionadas provienen de una lista de palabras objetivo diseñada específicamente para su tratamiento en rehabilitación (ver anexo C.4 Listado de palabras objetivo (DULCINEA)). Los resultados se valorarán considerando dos aspectos clave: el tiempo de procesamiento de cada etapa del pipeline y la calidad del contenido generado, evaluada a nivel de adecuación a las necesidades planteadas y de sincronización labial en las escenas con interacción entre los personajes.

El tiempo de ejecución total del pipeline para la generación de vídeos queda resumido en la Tabla 3.

En total tenemos un tiempo aproximado de 35 minutos por vídeo, lo que consume bastantes unidades computacionales de Google Colab Pro, limitando bastante la producción de vídeos al mes.

Etapa del pipeline	Tiempo de procesamiento	Tiempo total
Notebook 1		~22.5 min
Instalación de librerías		1.5 min
Instalación del repositorio Diffusers	1 min	
Generación de voces con Azure	0.5 min	
Cargar el modelo de baja resolución de Mochi1	7 min	7 min
Generación de escenarios	3.30 min/vídeo	~14 min
Notebook 2		~14 min
Instalación de dependencias	0.2 min	2 min
Instalación de librerías	1 min	
Clonación del repositorio Wav2Lip	0.5 min	
Adición de audios silenciosos a los contextos	0.2 min	
Concatenación para generar el vídeo final	0.1 min	
Sincronización de labios en inetarcciones	5-7 min/vídeo	~12 min
Tiempo total del sistema		**~36.5 min**

Tabla 3. Tabla secuencial con los tiempos de procesamiento para cada etapa del pipeline.

El proceso más demandante en términos de recursos y tiempo es la carga del modelo de baja resolución de Mochi 1 y la generación de cada escenario. El tiempo total necesario para completar el pipeline de generación de escenarios y personajes (ver Figura 12) en Google Colab Pro fue de aproximadamente 21 minutos, distribuidos entre la carga inicial del modelo Mochi 1 y la generación de los cuatro vídeos correspondientes al escenario. La carga del modelo, implementado con Diffusers en su variante de baja resolución, requirió aproximadamente 7 minutos debido al tamaño de los archivos preentrenados descargados desde Hugging Face y la inicialización del pipeline. La generación de cada vídeo individual, que representa las diferentes partes del escenario (contexto inicial, contexto final, interacción principal e interacción secundaria), tardó en promedio 3 minutos y 30 segundos por vídeo. Este tiempo incluye la interpretación de los prompts detallados que describen los atributos del entorno y los personajes. Las demás etapas del notebook 1, que incluye la instalación de librerías, la instalación del repositorio Diffusers y la ejecución de Azure para la generación de voces, tuvieron un tiempo de procesamiento total de entre 1 y 2 minutos.

En el notebook 2, el proceso de sincronización labial con Wav2Lip tomó entre 5 y 7 minutos por vídeo, lo que equivale a un total de aproximadamente 12 minutos para sincronizar las dos interacciones (principal y secundaria). Por otro lado, las operaciones adicionales, como la instalación de dependencias, la clonación del repositorio de Wav2Lip, la instalación de librerías requeridas antes de la sincronización, así como la adición de pistas de audio silenciosas a los contextos y la concatenación final de las cuatro partes del vídeo una vez generadas y sincronizadas las interacciones, tomaron aproximadamente 2 minutos en total.

4.1.1. MODELO DE BAJA RESOLUCIÓN DE MOCHI 1

El tiempo total del pipeline es de aproximadamente 35 minutos. En cuanto a la calidad del contenido generado, se puede observar que los vídeos presentan un nivel de detalle limitado. Los objetos en el entorno no se distinguen claramente ya que los algunos planos se superponen, como se puede visualizar especialmente en los contextos, y los rostros de los personajes carecen de definición.

Esta baja calidad afecta particularmente a la interacción secundaria, donde en este caso en concreto, el modelo no ha interpretado correctamente el prompt, pues se enfoca nuevamente al personaje principal en lugar de enfocar al personaje secundario en primer plano (ver Figura 15). Además, el pixelado en esta escena imposibilita que Wav2Lip reconozca las facciones del personaje y sincronice los labios con el audio. En el caso de la interacción principal, la sincronización labial es aceptable, pero la falta de realismo en los vídeos, atribuida a la precisión limitada del modelo de baja resolución, reduce significativamente la inmersión en la escena. Este aspecto evidencia que el sistema tiene un amplio margen de mejora, especialmente si se cuenta con mayores recursos computacionales para cargar el modelo de Mochi 1 de alta resolución, lo que permitiría generar vídeos con mayor definición.

Figura 15. Ejemplo de una escena generada mediante el modelo de baja resolución de Mochi 1, que se desarrolla en la consulta de un médico. El paciente (personaje principal) acude al doctor, le cuenta cómo se siente y el doctor le responde. La escena termina con el paciente saliendo de la sala, aunque en este caso en particular, el vídeo no se adhiere exactamente al prompt utilizado.

4.1.2. VÍDEOS GENERADOS ONLINE E IMPORTADOS EN EL PIPELINE

La calidad de los vídeos generados con el modelo de baja resolución es limitada, mostrando poca definición en los detalles y dificultando la percepción precisa de los escenarios y personajes. Sin embargo, al trabajar con vídeos descargados directamente del playground de Mochi 1 o, en el mejor de los casos, utilizando el modelo de alta resolución en un entorno técnico más robusto, la calidad mejora significativamente. Los detalles de los vídeos están mejor definidos y los personajes se visualizan con mayor nitidez, proporcionando una representación más realista de las escenas. Aun así, la continuidad entre las diferentes partes del vídeo final todavía presenta áreas de mejora, no solo en la consistencia de los escenarios, sino también en la coherencia de los personajes que aparecen a lo largo de las escenas. Además, las acciones y movimientos de algunas de las personas que intervienen resultan algo forzadas, restando realismo. Las voces generadas, aunque son funcionales, todavía están lejos de simular una entonación completamente natural, ya que en ocasiones suenan un poco robóticas. Cabe destacar que no se ha explorado en profundidad el ajuste de parámetros de prosodia disponibles en Azure Speech y comentado en el apartado 3.1.2, lo cual podría ser una línea futura interesante para mejorar este aspecto. En cuanto a la sincronización labial, no es siempre perfecta y muestra variaciones dependiendo del vídeo. En algunos casos, aparece un halo cuadrado poco perceptible alrededor de la cara del personaje al realizar la sincronización. Aunque no molesta visualmente, sí evidencia que se está llevando a cabo un proceso de ajuste técnico.

A continuación, se pueden observar los siguientes vídeos generados en las figuras: Figura 16, Figura 17, Figura 18, Figura 19, Figura 20, Figura 21.

Figura 16. https://www.youtube.com/watch?v=LBcx61Iz1sM Ejemplo de una escena que se desarrolla en una cafetería. Un hombre de mediana edad (personaje principal) entra, pide un café, y se sienta en una mesa con su mujer. Se percibe un gran nivel de detalle en el café caliente que sostiene la camarera. Las voces son naturales y su sincronización es buena. Sin embargo, existen inconsistencias visuales como el cuello de la camisa del hombre.

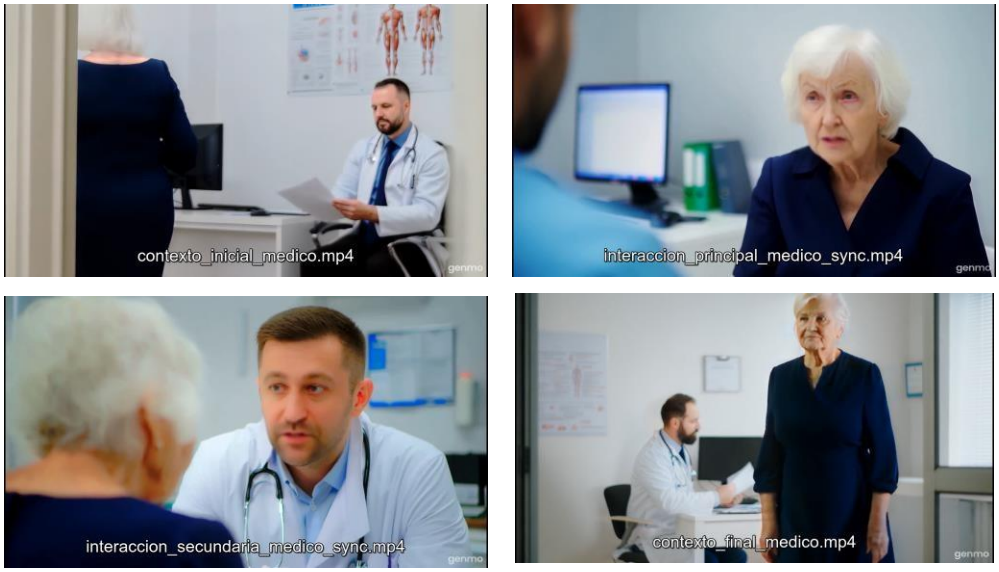

Figura 17. https://www.youtube.com/watch?v=J_ZdiQYRlFY Ejemplo de una escena que se desarrolla en la consulta de un médico. Una mujer mayor (personaje principal) entra en la consulta, se sienta y le cuenta al doctor cómo se encuentra. El doctor le responde y la mujer se dirige hacia la salida. Las voces suenan naturales y la sincronización es buena. Sin embargo, en el contexto final de este vídeo se puede observar que la postura y el movimiento de la mujer es bastante forzado. Además, el doctor (personaje secundario) presenta diferencias bastante notables a lo largo de la secuencia de las escenas.

Figura 18. https://www.youtube.com/watch?v=L3f_8ocWF9A Ejemplo de una escena que se desarrolla en una calle residencial. Una mujer de mediana edad (personaje principal) se dirige a coger un taxi. Una vez dentro del coche le indica al taxista la dirección a la que necesita ir y el taxista le responde. En la escena final se ve a la mujer llegando a casa. En el contexto inicial, la señora realiza un movimiento extraño a la hora de acercarse al taxi, demostrando que hay ciertas acciones que a la IA le cuesta más representar. Las voces y la sincronización son correctas.

Figura 19. https://www.youtube.com/watch?v=5ApoutbdCEs Ejemplo de una escena que se desarrolla en casa. El hombre (personaje principal) está en el salón sentado en el sofá y le propone a su mujer salir a dar un paseo. El contexto final les muestra andando de la mano por el parque. En este caso, los vídeos se adhieren bastante bien a los prompts especificados y el movimiento es menos forzado que en otras ocasiones. Las voces generadas y la sincronización labial son buenas.

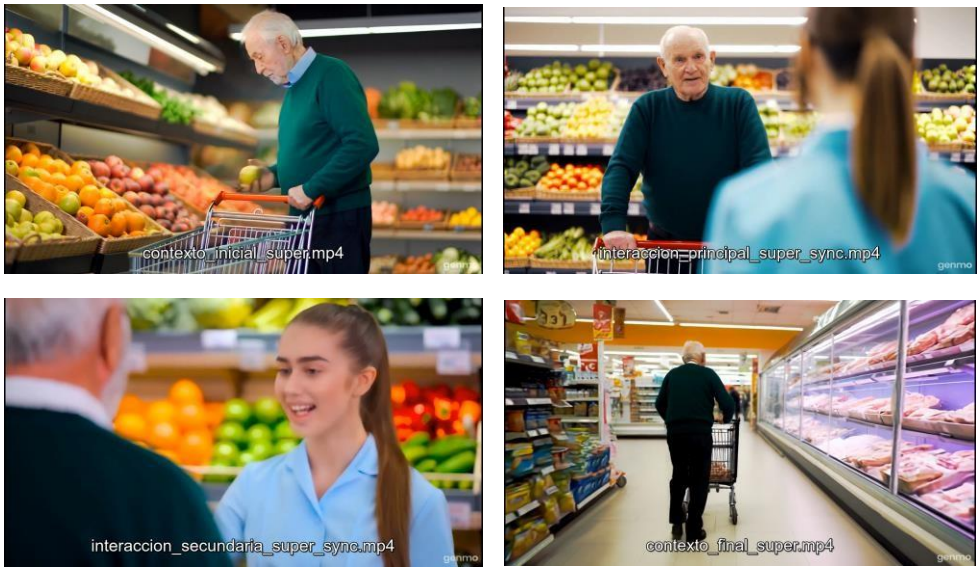

Figura 20. https://www.youtube.com/watch?v=AtxhjMDYbRE Ejemplo de una escena que se desarrolla en un supermercado. Un señor mayor (personaje principal) está en la sección de frutas y verduras, y se dirige a una dependienta hacerle una consulta. Ella le responde y él se dirige hacia esa dirección empujando su carrito de la compra. En este caso, el color del carro es distinto en las escenas y el hombre también presenta ligeras diferencias físicas. En cuanto a las voces y la sincronización labial, la voz de la mujer se escucha un poco robótica y los labios del hombre no están del todo sincronizados con el audio.

Figura 21. https://www.youtube.com/watch?v=OOmQZEWHzDA Ejemplo de una escena que se desarrolla en casa. Una mujer de mediana edad (personaje principal) está en el salón con su hija que le está enseñando algo que ha escrito. La hija se sienta al lado de su madre y la madre le da la enhorabuena. La hija le responde y se levantan para darse un abrazo. La madre presenta ligeras diferencias en el pelo. Las voces son adecuadas y la sincronización labial es buena. Además, durante el abrazo, los movimientos corporales son bastantes realistas.

En resumen, los resultados obtenidos reflejan tanto el potencial como las limitaciones del sistema desarrollado. Aunque el pipeline logra automatizar de manera significativa la generación de vídeos personalizados y contextualizados, los resultados actuales evidencian la necesidad de mayores recursos computacionales para alcanzar una calidad de contenido que cumpla con los estándares clínicos. La combinación de herramientas como Mochi 1, Azure Speech y Wav2Lip demuestra ser una opción aceptable para simular escenarios inmersivos, pero aspectos como la sincronización labial, la fluidez de las voces y la continuidad de las escenas todavía necesitan mejoras. Estos resultados destacan la importancia de seguir trabajando en la optimización técnica y en la integración de modelos más avanzados que permitan aumentar tanto la resolución de los vídeos como la personalización en futuros desarrollos.

5. CONCLUSIONES Y LÍNEAS FUTURAS

5.1. CONCLUSIONES

El desarrollo de este proyecto ha demostrado el potencial de la IA generativa como herramienta aplicada en el ámbito de la salud, en concreto en la rehabilitación de pacientes con afasia. La integración de modelos avanzados como Mochi 1 para la generación de vídeos, Azure Speech para la creación de audios, y Wav2Lip para la sincronización labial, ha permitido el diseño de un sistema automatizado capaz de generar escenas personalizadas y contextualizadas que simulan interacciones reales. Su enfoque innovador radica en la inmersión de los pacientes en escenas cotidianas, donde pueden verse reflejados y participar activamente. Cabe destacar el alto nivel de personalización que ofrece la IA generativa, ya que los clínicos tienen la posibilidad de decidir sobre qué escenarios trabajar, qué personajes incluir y qué frases practicar, evitando las limitaciones de ceñirse a material pregrabado o estandarizado. Esta capacidad de personalización no solo refuerza la inmersión y motivación del paciente, sino que también permite adaptar la terapia a sus necesidades específicas y a los objetivos definidos por el terapeuta.

Más que una solución definitiva, este TFG debe entenderse como un punto de partida o un prototipo inicial, que pueda ser desarrollado y mejorado en un futuro con el objetivo de ser aplicado como técnica de rehabilitación complementaria a las actuales. Esto se debe tanto a las limitaciones computacionales como a las inherentes al uso de modelos generativos avanzados.

En cuanto a las limitaciones propias de las herramientas de IA utilizadas, se puede distinguir una ligera desincronización entre audio y vídeo a la hora de sincronizar los labios con Wav2Lip. Respecto a las voces generadas con la versión gratuita de la API de Azure Speech, podemos concluir que, las opciones en español son limitadas tanto en cantidad como en estilos disponibles, dificultando la integración de emociones en los diálogos. Sin embargo, se han expuesto ciertos parámetros ajustables, como el tono, la velocidad y la prosodia, que permiten modular las voces adecuándolas a las necesidades clínicas. Acerca de la generación de avatares y escenarios, se ha podido observar en los resultados la dificultad de escribir prompts consistentes que garanticen la coherencia y continuidad de las escenas.

Además, en relación con el modelo de generación de escenarios Mochi 1, se pueden destacar los altos requisitos computacionales que suponen una barrera significativa en la elaboración de este proyecto. Debido a las limitaciones de capacidad de cómputo, solo fue posible cargar el modelo de baja resolución de Mochi 1 en Google Colab Pro, lo que se vio reflejado en vídeos de muy baja calidad. En algunos casos, esta baja calidad ha dificultado que herramientas como Wav2Lip detectasen correctamente los rostros en los vídeos, impidiendo realizar una sincronización labial precisa. De esta manera, se ha optado por generar demos con vídeos descargados directamente del playground online de Mochi 1, los cuales fueron posteriormente importados al sistema de generación automático de escenas, donde se generaron las voces y se realizó la sincronización labial. Este enfoque se realizó con el objetivo de visualizar cómo podrían ser los resultados finales si se contase con los recursos técnicos necesarios para generar vídeos de alta calidad de manera directa.

Es importante destacar que los resultados obtenidos en este proyecto se lograron utilizando herramientas y recursos de bajo presupuesto. Esto implica que, con una mínima inversión económica, los resultados podrían ser significativamente mejores, especialmente en términos de calidad visual y sincronización. Herramientas avanzadas o planes de pago permitirían aprovechar al máximo las funcionalidades técnicas de los modelos utilizados, resolviendo muchas de las limitaciones actuales. Asimismo, la IA generativa es un campo novedoso en continuo desarrollo, por lo que es altamente probable que surjan nuevas tecnologías y modelos generativos con mejores capacidades y funcionalidades.

Finalmente, este proyecto representa un punto de partida significativo hacia la incorporación de IA generativa en entornos clínicos. Demuestra su capacidad para transformar la forma en que se aborda la rehabilitación del lenguaje actualmente y expone una propuesta complementaria, innovadora, inmersiva y personalizada con el fin de mejorar la experiencia terapéutica de los pacientes. Sin embargo, aún quedan muchas mejoras que incluir en desarrollos posteriores, así como limitaciones a las que hacer

frente, como la optimización de recursos computacionales, la integración en sistemas hospitalarios y el desarrollo de funcionalidades adicionales que amplíen su alcance.

5.2. LÍNEAS FUTURAS

Con el fin de continuar con esta línea de investigación y mejorar los resultados obtenidos, se proponen las siguientes líneas futuras clasificadas según el campo de mejora.

5.2.1. MEJORAS TÉCNICAS Y OPTIMIZACIÓN DE RECURSOS COMPUTACIONALES

- **Planes de pago**: El único coste de este proyecto viene dado por la suscripción mensual de Google Colab Pro. Considerar planes de pago accesibles de herramientas de IA y elaborar presupuestos para maximizar las funcionalidades técnicas podría ser de utilidad en futuras mejoras de calidad y funcionamiento.

- **Optimización de recursos computacionales**: Intentar reducir los requerimientos técnicos evaluando nuevas tecnologías y versiones optimizadas de los modelos. Si no fuese viable, intentar contratar un plan de pago con mejores recursos donde se pueda ejecutar el modelo de alta resolución de Mochi 1.

- **Evaluar herramientas de IA más eficientes**: Valorar modelos como Diff2Lip que han demostrado un gran rendimiento en sincronización labial podría optimizar resultados en futuros desarrollos [55]. Asimismo, explorar nuevos modelos generativos de escenarios o voces y su integración en un pipeline automatizado.

- **Mejoras en la arquitectura del sistema**: Automatizar aún más el flujo de trabajo para minimizar el tiempo de configuración y generación de las escenas, y agilizar el proceso para usuarios con menos experiencia técnica. Esto incluiría la implementación de scripts o configuraciones predefinidas que permitan establecer el entorno de ejecución de manera automática, evitando que el usuario tenga que introducir manualmente las líneas de código necesarias para instalar librerías, cargar modelos o configurar las dependencias.

5.2.2. MEJORAS EN LA CALIDAD DEL CONTENIDO GENERADO

- **Voces e imágenes personalizadas**: Una vez superadas las limitaciones técnicas actuales, sería posible explorar la inclusión de voces e imágenes de los propios pacientes en el proyecto, permitiendo que la IA genere contenido audiovisual a partir de estos. Esta idea sería interesante en el caso de necesitar vídeos aún más personalizados. Sin embargo, su implementación no resulta tan prioritaria como la mejora de las limitaciones técnicas relacionadas con la calidad de los resultados actuales.

- **Optimizar el diseño de los prompts**: Mejorar los prompts utilizados en la generación de vídeos para asegurar la consistencia entre las escenas. Diseñar mejores instrucciones que sigan manteniendo un alto nivel de detalle para que el modelo pueda interpretarlo, pero evitando construcciones ambiguas o excesivamente largas que puedan confundir al generador y desperdiciar los recursos disponibles.

- **Escenas con más interacciones:** Actualmente, en la generación de escenarios, solo se incluye una interacción por parte de cada personaje. Añadir más vídeos donde la escena no se limite a simplemente una frase del personaje principal y la respuesta del personaje secundario, si no una

conversación con más interacciones por parte de estos incluso con más personajes, podría ser una línea de mejora muy interesante para el futuro.

- **Añadir ruido de fondo:** En este TFG, el único audio que se incluye corresponde a las voces generadas para las interacciones. Sin embargo, como mejora a implementar en el futuro, sería interesante añadir ruido de fondo en los contextos para simular ambientes más realistas, como el bullicio de una cafetería o el ruido de una calle concurrida. En las interacciones, además del diálogo, el ruido de fondo podría añadirse como elemento de dificultad, ya que algunos pacientes que tuve la oportunidad de conocer mencionaban lo complicado que les parecía mantener la atención en su conversación en ambientes donde hubiera otras personas hablando. Incluir esta complejidad en las terapias podría ser un recurso para mejorar la capacidad de concentración y de comprensión auditiva de los pacientes.

- **Base de datos de atributos**: Actualmente, es necesario escribir los atributos de entrada manualmente a la hora de generar escenarios y voces con los archivos mochi.py y azure.py respectivamente. Una posible mejora consistiría en crear una BBDD dinámica donde estén registrados escenarios y personajes con sus características principales. De esta manera, los terapeutas podrían añadir nuevos atributos y seleccionar fácilmente entre opciones ya disponibles sin necesidad de escribir prompts manualmente cada vez que se quiera generar una nueva escena. También podría crearse un archivo JSON externo donde se almacenen configuraciones de escenarios predefinidos, simplificando la creación y personalización de contenido en tiempo real.

- **Contextos específicos**: Cambiar el enfoque de palabras objetivo, visto en proyectos como Dulcinea, a situaciones específicas como “estar en una cafetería,” y a partir de este ahí trabajar todas las posibles frases que se podrían utilizar en este contexto.

- **Subtítulos**: Añadir subtítulos a los vídeos generados, lo que no solo facilitaría la comprensión del contenido, sino que también podría adaptarse a pacientes con necesidades adicionales, como hipoacusia.

5.2.3. EXPANSIÓN Y ESCALABILIDAD DEL SISTEMA

- **Extrapolación internacional:** Adaptar el sistema para su uso fuera de España aprovechando la gran cantidad de voces disponibles en diferentes idiomas.

- **Software o soporte hospitalario:** Diseñar software específico para integrar este sistema en hospitales, con interfaces intuitivas que permitan a los terapeutas gestionar las sesiones y adaptar los escenarios de manera ágil.

6. BIBLIOGRAFÍA

[1] Real Academia Española. (s.f.). Diccionario de la lengua española (23.ª ed.). Recuperado de https://dle.rae.es/

[2] González Victoriano, R., & Hornauer-Hughes, A. (2014). Afasia: una perspectiva clínica. Revista de Neurología, 25(291-308).

[3] Real Academia Nacional de Medicina. (s.f.). Diccionario terminológico de la medicina española. Madrid: Real Academia Nacional de Medicina.

[4] Afasia.org. (s.f.). Afasia: información, recursos y apoyo. Recuperado de https://afasia.org/

[5] Sociedad Española de Neurología. (s.f.). Afasia y sus implicaciones clínicas. Recuperado de https://www.sen.es/saladeprensa/pdf/Link223.pdf

[6] National Aphasia Association. (2024). What is aphasia? Recuperado de https://aphasia.org/what-is-aphasia/

[7] Biblioteca Nacional de Medicina de EE. UU. (2020). Afasia. Recuperado de https://medlineplus.gov/spanish/aphasia.html

[8] Siniscalchi, A. (2022). Use of stroke scales in clinical practice: Current concepts. Turkish Journal of Emergency Medicine, 22(3), 119-124.

[9] National Institute on Deafness and Other Communication Disorders (NIDCD). (s.f.). Afasia. Recuperado de https://www.nidcd.nih.gov/es/espanol/afasia

[10] Benedetti, F., & P. Z. (2018). Current evidence on transcranial magnetic stimulation in aphasia. Neurología (English Edition), 33(9), 591-598.

[11] De la Serna, J. M. (s.f.). Modelo de Wernicke-Geschwind (1960s): Hito en neurolingüística. Recuperado de https://juanmoisesdelaserna.es/ramas-de-las-neurociencias-descubre-todas-las-areas-y-especialidades/neurolinguistica-bases-cerebrales-del-lenguaje/evolucion-de-la-neurolinguistica-timeline/modelo-de-wernicke-geschwind-1960s-hito-en-neurolinguistica/

[12] ISEP. (s.f.). Afasia, uno de los ámbitos más frecuentes de la logopedia. Recuperado de https://www.isep.es/actualidad/afasia-uno-de-los-ambitos-mas-frecuentes-de-la-logopedia/

[13] Pyun, S. B., Sohn, H. J., Jung, J. H., Nam, K., & Kim, M. (2015). Community integration and quality of life in aphasia after stroke. Yonsei Medical Journal, 56(6), 1694–1702.

[14] Hilari, K. (2011). The impact of stroke: Are people with aphasia different to those without? Disability and Rehabilitation, 33(3), 211–218.

[15] Pillay, B. F. S. M., & C. P. (2017). Social participation in working-age adults with aphasia: An updated systematic review. Topics in Stroke Rehabilitation, 24(8), 627–639.

[16] Clínica Nodos. (s.f.). El papel del logopeda en pacientes con afasia. Recuperado de https://clinicanodos.es/el-papel-del-logopeda-en-pacientes-con-afasia/

[17] Meinzer, R. B., Meinzer, M. E., & D. W. (2004). Intensive language training enhances brain plasticity in chronic aphasia. BMC Biology, 2, 20.

[18] Peitz, D., Schumann-Werner, B., Hussmann, K. et al. (2024). Success rates of intensive aphasia therapy: real-world data from 448 patients between 2003 and 2020. Journal of Neurology, 271, 7169–7183.

[19] Trivium. (s.f.). Neurorehabilitación y la afasia. Recuperado de https://trivium.cat/neurorehabilitacion-la-afasia/

[20] Brady, M. C., Kelly, H., Godwin, J., Enderby, P., & Campbell, P. (2016). Speech and language therapy for aphasia following stroke. Cochrane Database of Systematic Reviews, 2016(6), CD000425.

[21] My Wellness Hub. (s.f.). Best speech therapy activities for aphasia in adults. Recuperado de https://www.mywellnesshub.in/blog/best-speech-therapy-activities-for-aphasia-in-adults/

[22] Delgado Santos, C. I. (2013). Cuaderno de apoyo a la comunicación en el entorno sanitario: Personas con afasia. CEAPAT-IMSERSO.

[23] National Institutes of Health. (2020, julio). Desglose de la comunicación: Cómo la afasia afecta el lenguaje. Recuperado de https://salud.nih.gov/recursos-de-salud/nih-noticias-de-salud/desglose-de-la-comunicacion

[24] Mayo Clinic. (s.f.). Afasia - Diagnosis and treatment. Recuperado de https://www.mayoclinic.org/diseases-conditions/aphasia/diagnosis-treatment/drc-20369523

[25] Pulvermüller, F., Neininger, B., Elbert, T., Mohr, B., Rockstroh, B., Koebbel, P., & Taub, E. (2001). Constraint-induced therapy of chronic aphasia after stroke. Stroke: A Journal of Cerebral Circulation, 32(7), 1621-1626.

[26] Cao, Y. F., Liu, G. H., & Zhang, J. (2024). A computer-aid speech rehabilitation system with mirrored vídeo generating. Technology and Health Care, 32, S543-S553.

[27] Arya, K. N., & Pandian, S. (2014). Inadvertent recovery in communication deficits following the upper limb mirror therapy in stroke: A case report. Journal of Bodywork and Movement Therapies, 18(4), 566-568.

[28] Fuentes, B., de la Fuente-Gómez, L., Sempere-Iborra, C., Delgado-Fernández, C., Tarifa-Rodríguez, A., Alonso de Leciñana, M., de Celis-Ruiz, E., Gutiérrez-Zúñiga, R., López-Tàppero, J., Martín Alonso, M., Pastor-Yborra, S., Rigual, R., Ruiz-Ares, G., Rodríguez-Pardo, J., Virués-Ortega, J., Borobia, A. M., Blanco, P., & Bueno-Guerra, N. (2022). DUbbing Language-therapy CINEMA-based in Aphasia post-Stroke (DULCINEA): study protocol for a randomized crossover pilot trial. Trials, 23(1), 21. https://doi.org/10.1186/s13063-021-05956-5

[29] Fuentes, B., Jordi-Perea, P., Sempere-Iborra, C., Tarifa-Rodríguez, A., de Celis-Ruiz, E., Martín Alonso, M., Ruiz-Ares, G., Rigual, R., Rodríguez-Pardo, J., Alonso-López, E., Alonso de Leciñana, M., Virués-Ortega, J., Borobia, A. M., Jiménez-González, M., Martínez-Balaguer, M., Blanco, P., & Bueno, N. (2024). Dubbing language-therapy CINEMA-based in aphasia post-stroke (DULCINEA): A feasibility randomized crossover controlled trial. Digital Health, 10, 20552076241288311. https://doi.org/10.1177/20552076241288311

[30] IBM. (s.f.). Artificial Intelligence. Recuperado de https://www.ibm.com/mx-es/topics/artificial-intelligence

[31] Digital 55. (s.f.). Inteligencia artificial, machine learning y deep learning. Recuperado de https://digital55.com/blog/inteligencia-artificial-machine-learning-y-deep-learning/

[32] Hugging Face. (s.f.). Using diffusers for text-to-vídeo generation. Recuperado de https://huggingface.co/docs/diffusers/main/en/using-diffusers/text-img2vid

[33] Runway ML. (s.f.). Pricing. Recuperado de https://runwayml.com/pricing

[34] Runway ML. (s.f.). Billing. Recuperado de https://docs.dev.runwayml.com/usage/billing/

[35] Stability AI. (s.f.). Remove background from image using the API. Recuperado de https://platform.stability.ai/docs/api-reference#tag/Image-to-Vídeo

[36] Hugging Face. (s.f.). Using diffusers for singular value decomposition (SVD). Recuperado de https://huggingface.co/docs/diffusers/main/en/using-diffusers/svd

[37] Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., & Rombach, R. (2023). Stable vídeo diffusion: Scaling latent vídeo diffusion models to large datasets. Recuperado de https://static1.squarespace.com/static/6213c340453c3f502425776e/t/655ce779b9d47d342a93c890/1733935148453/stable_vídeo_diffusion.pdf

[38] Heygen. (s.f.). Heygen: AI-powered vídeo creation platform. Recuperado de https://www.heygen.com/?sid=rewardful&via=m2&msclkid=6fbfbb5b8d5e1fd9edc6dc3d7b3baf23

[39] Synthesia. (s.f.). Opciones de precios. Recuperado de https://www.synthesia.io/es/opciones-de-precios

[40] Synthesia. (s.f.). Introduction to the Synthesia API. Recuperado de https://docs.synthesia.io/reference/introduction

[41] HailuoAI. (s.f.). Subscribe to Hailuo AI. Recuperado de https://hailuoai.vídeo/subscribe

[42] Minimaxi. (s.f.). Platform Introduction. Recuperado de https://intl.minimaxi.com/document/platform%20introduction?key=66701c8e1d57f38758d58198

[43] D-ID. (s.f.). Pricing Studio. Recuperado de https://www.d-id.com/pricing/studio/

[44] D-ID. (s.f.). Interview answering: Developer's questions about D-ID's API. Recuperado de https://www.d-id.com/blog/interview-answering-developers-questions-about-d-ids-api/

[45] Pikart AI. (s.f.). API. Recuperado de https://pikartai.com/api/

[46] Genmo AI. (s.f.). Blog. Recuperado de https://www.genmo.ai/blog

[47] Mochi AI. (s.f.). Mochi AI Homepage. Recuperado de https://mochi-1.ai/

[48] Mochi AI. (s.f.). Pricing. Recuperado de https://mochi1ai.com/pricing

[49] DeepMind. (s.f.). VEO-2. Recuperado de https://deepmind.google/technologies/veo/veo-2/

[50] Microsoft. (s.f.). Azure AI Services Speech Service. Recuperado de https://learn.microsoft.com/es-es/azure/ai-services/speech-service/overview

[51] Eleven Labs. (s.f.). Eleven Labs – AI Voice Generation. Recuperado de https://elevenlabs.io/

[52] Rudrabha. (s.f.). Wav2Lip. Recuperado de https://github.com/Rudrabha/Wav2Lip

[53] Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. V. (2020). A lip sync expert is all you need for speech to lip generation in the wild. Proceedings of the 28th ACM International Conference on Multimedia (MM ’20), 1–10. https://doi.org/10.1145/3394171.3413532

[54] Google. (s.f.). Precios de los servicios pagados de Colab. Recuperado de https://colab.research.google.com/signup

[55] Mukhopadhyay, S., Suri, S., Gadde, R. T., & Shrivastava, A. (2024). Diff2Lip: Audio conditioned diffusion models for lip-synchronization. 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 5280–5290. https://doi.org/10.1109/WACV57701.2024.00521

ANEXO A: ASPECTOS ÉTICOS, ECONÓMICOS, SOCIALES Y AMBIENTALES

A.1 INTRODUCCIÓN

Este TFG se desarrolla bajo el contexto de un proyecto previo que exploraba una nueva técnica de rehabilitación en pacientes con afasia que consiste en practicar la enunciación de frases mediante el doblaje de escenas contextualizadas. Con el objetivo de poder desvincularse y dejar de depender de contenido pregrabado como escenas de series de televisión, este TFG muestra un sistema de generación de vídeos mediante IA generativa. Esto aporta un gran nivel de personalización en las terapias, pues se puede practicar la situación y las frases que más convengan para cada paciente, siendo una solución motivadora y con mucho potencial a desarrollar.

A.2 DESCRIPCIÓN DE IMPACTOS RELEVANTES RELACIONADOS CON EL PROYECTO

- **Impacto social**: La afasia post-ictus es una condición que afecta a la calidad de vida de los pacientes, limitando su capacidad de comunicarse y socializar. Según el consenso de pacientes, cuidadores y profesionales de la salud, ayudar a las personas a recuperarse de la afasia se encuentra entre las 10 prioridades principales de investigación relacionadas con la vida después de un ictus [29]. Este proyecto busca ofrecer una técnica terapéutica innovadora que podría mejorar la inclusión de manera más sencilla para los pacientes, aumentando su motivación ofreciéndoles la oportunidad de practicar con escenarios personalizados, según las dificultades de cada uno.

- **Impacto económico**: El ictus representa el 70% de los ingresos neurológicos que se producen en España y es responsable de entre el 3 y el 6% del gasto total sanitario [5]. Estos datos sugieren la cantidad de personas que podrían beneficiarse de la implementación de terapias como la descrita en este TFG para la rehabilitación de una de sus secuelas más limitantes, la afasia. Además, estos datos no solo reflejan la gravedad de esta condición, sino también los elevados costes asociados, que van más allá de la hospitalización inicial e incluyen la rehabilitación a largo plazo de los pacientes. Este TFG se presenta como una posible solución para aliviar esta carga económica relacionada con el tratamiento de los pacientes con afasia.

- **Impacto ético**: La ética en el uso de la IA es un aspecto crucial en este proyecto. Garantizar la privacidad de los pacientes y el uso responsable de sus datos es prioritario, especialmente si en el futuro se integra la opción de incluir las voces o imágenes de los propios usuarios en las terapias. La IA y la tecnología, utilizada de manera responsable, tiene el potencial de favorecer significativamente la calidad de vida de las personas. Además, este proyecto promueve un enfoque inclusivo, asegurando que las herramientas sean accesibles y equitativas, respetando los principios éticos de justicia y no discriminación en el acceso a la tecnología.

- **Impacto medioambiental**: La ejecución de este sistema en plataformas como Google Colab Pro permite evitar la adquisición de hardware personal. Si bien es cierto que el consumo energético y de recursos computacionales proporcionan una huella ambiental considerable.

A.3 ANÁLISIS DETALLADO DE ALGUNO DE LOS PRINCIPALES IMPACTOS

Uno de los impactos más relevantes está relacionado con la accesibilidad social y económica del sistema. Este proyecto utiliza un modelo de bajo presupuesto con herramientas como Google Colab Pro y librerías de código abierto, lo que reduce las barreras económicas y permite que las terapias puedan

implementarse en una amplia gama de contextos, desde grandes hospitales hasta clínicas pequeñas. Sin embargo, este enfoque también refleja limitaciones en términos de calidad de los vídeos generados debido a restricciones computacionales, lo que destaca la necesidad de realizar inversiones adicionales en hardware o suscripciones avanzadas en el futuro.

A.4 CONCLUSIONES

Desde una perspectiva ética, social y económica, este proyecto aporta una solución innovadora para mejorar las terapias de rehabilitación en pacientes con afasia. La inclusión de criterios de sostenibilidad, como el uso de tecnologías de código abierto y recursos computacionales accesibles, ha permitido reducir los costes y fomentar un enfoque inclusivo. No obstante, la optimización de los recursos computacionales y la mejora en la calidad del contenido son aspectos a continuar desarrollando maximizar su impacto social y clínico en un futuro.

ANEXO B: PRESUPUESTO ECONÓMICO

COSTE DE MANO DE OBRA (coste directo)	Horas	Precio/hora	Total
	360	20 €	**7.200,00 €**

COSTE DE RECURSOS MATERIALES (coste directo)	Precio de compra	Uso en meses	Amortización (en años)	Total
Ordenador personal (Software incluido)	1.500,00 €	3	4	93,75 €
Suscripción mensual Google Colab Pro	11,19 €	3		33,57 €
Total recursos materiales				**127,32 €**

TOTAL COSTES DIRECTOS	**7.327,32 €**

GASTOS GENERALES (costes indirectos)	15%	sobre CD	**1.099,10 €**
BENEFICIO INDUSTRIAL	6%	sobre CD+CI	**439,64 €**

MATERIAL FUNGIBLE

No aplicable	**00,00 €**
No aplicable	**00,00 €**

SUBTOTAL PRESUPUESTO		**8.866,06 €**
IVA APLICABLE	21%	**1.861,87 €**

TOTAL PRESUPUESTO	**10.727,93 €**

Tabla 4. Para la realización de este TFG se han dedicado 360 horas y se ha estimado un precio de 20 €/ hora.

Como recursos materiales se ha incluido el uso durante 3 meses del ordenador personal y la suscripción mensual a Google Colab Pro para la utilización de sus servicios computacionales. No existe material fungible que incluir en este proyecto.

El presupuesto total asciende a 10.727,93.

ANEXO C: MATERIAL ADICIONAL

C.1 NIHSS

National Institutes of Health Stroke Scale.

Escala de Ictus del National Institute of Health (NIHSS)

1.a. Nivel de conciencia	Alerta	0
	No alerta (mínimos estímulos verbales)	1
	No alerta (estímulos repetidos o dolorosos)	2
	Respuestas reflejas	3
1.b. Preguntas	Ambas respuestas correctas	0
¿En qué mes estamos?	Una respuesta correcta (o disartria	1
¿Qué edad tiene?	Ninguna respuesta correcta (o afasia)	2
1.b. Órdenes motoras	Ambas órdenes correctas	0
1. Cierre los ojos	Una orden correcta	1
2. Abra y cierre la mano	Ninguna orden correcta	2
2. Mirada conjugada (horizontal)	Normal	0
	Parálisis parcial de la mirada	1
	Desviación forzada de la mirada	2
3. Campo visual	Normal	0
	Hemianopsia Parcial	1
	Hemianopsia Completa	2
	Ceguera	3
4. Paresia facial	Movilidad Normal	0
	Paresia menor	1
	Paresia parcial	2
	Parálisis completa de la hemicara	3
5. Miembro superior derecho / miembro superior izquierdo	No caída del miembro	0/0
	Caída en menos de 10 segundos	1/1
	Esfuerzo contra la gravedad	2/2
	Movimiento en el Plano horizontal	3/3
	No movimiento	4/4
6. Miembro inferior derecho / miembro inferior izquierdo	No caída del miembro	0/0
	Caída en menos de 5 segundos	1/1
	Esfuerzo contra la gravedad	2/2
	Movimiento en el Plano horizontal	3/3
	No movimiento	4/4
7. Ataxia de Miembros	Ausente	0
	Presente en 1 extremidad	1
	En 2 o más extremidades	2
8. Exploración Sensitiva	Normal	0
	Perdida entre ligera a moderada	1
	Perdida entre grave y total	2
9. Lenguaje	Normal	0
	Afasia ligera a moderada	1
	Afasia grave	2
	Afasia global	3
10. Disartria	Normal	0
	Ligera a moderada	1
	Grave a anartria	2
11. Extinción e Inatención (negligencia)	Normal	0
	Extinción parcial	1
	Extinción completa	2
Total (máximo 42)		

C.2 CIQ

Cuestionario de Integración en la Comunidad.

Community Integration Questionnaire

Name:_____ Date:_____

Home Integration	Answer (circle one)	Score
1. Who usually does shopping for groceries or other necessities in your household?	Yourself alone (2) Yourself and someone else (1) Someone else (0)	
2. Who usually prepares meals in your household?	Yourself alone (2) Yourself and someone else (1) Someone else (0)	
3. In your home who usually does normal everyday housework?	Yourself alone (2) Yourself and someone else (1) Someone else (0)	
4. Who usually cares for the children in your home?	Yourself alone (2) Yourself and someone else (1) Someone else (0) Not applicable (score is the average of 1,2,3 and 5)	
5. Who usually plans social arrangements such as get-togethers with family and friends?	Yourself alone (2) Yourself and someone else (1) Someone else (0)	
Home Integration Total Score	Add the above scores together	
Social Integration		
6. Who usually looks after your personal finances such as banking or paying bills?	Yourself alone (2) Yourself and someone else (1) Someone else (0)	
Can you tell me approximately how many times a month you now usually participate in the following activities outside your home?		
7. Shopping	5 or more (2) 1 – 4 times (1) Never (0)	
8. Leisure activities such as movies, sports, restaurants	5 or more (2) 1 – 4 times (1) Never (0)	
9. Visiting friends or relatives	5 or more (2) 1 – 4 times (1) Never (0)	

C.3 BDAE

Examen Boston de Diagnóstico de Afasia.

TEST DE BOSTON PARA EL DIAGNOSTICO DE LA AFASIA
FORMATO ESTANDAR*

1. Identificación del Paciente.

Nombre: _____ **Edad**: _____
N° de Ficha: _____ **Fecha de Nacimiento**: _____
Antecedentes Clínicos: _____
_____
_____

2. Evaluación.

Fecha de Inicio: _____ **Fecha de Término**: _____

Escala de Severidad de la Afasia.	/5	0	10	20	30	40	50	60	70	80	90	100
		0	0	1	1	1	2	3	3	3	4	5

Perfil de Características del Habla.		0	10	20	30	40	50	60	70	80	90	100
Agilidad Articulatoria.	/7	1	2	3	3	4	5	6	6	7	7	7
Longitud de la Frase.	/7	1	2	4	6	7	7	7	7	7	7	7
Forma Gramatical.	/7	1	2	3	4	5	5	6	6	7	7	7
Línea Melódica.	/7	1	2	3	5	5	6	6	7	7	7	7
Parafasias.	/7	1	2	2	3	4	5	6	6	7	7	7

		1	2	3	4	5	6	7
1.	AGILIDAD ARTICULATORIA. Facilidad a nivel fonémico y silábico.	Incapaz de formar los sonidos del habla.			A veces torpe o esforzada.			Nunca defectuosa.
2.	LONGITUD DE LA FRASE. En habla espontánea la más larga interrumpida de palabras.	1 palabra.			4 palabras.			7 palabras.
3.	FORMA GRAMATICAL. Variedad de construcciones gramaticales; uso de morfemas gramaticales.	Sin agrupamientos sintácticos de palabras.			Formas simplificadas o incompletas; omisiones de morfemas gramaticales.			Rango normal de sintaxis, facilidad normal con las palabras gramaticales.
4.	LINEA MELÓDICA (PROSODIA).	Palabra por palabra o habla aprosódica.			Entonación de oraciones limitada a frases cortas.			Melodía normal.
5.	PARAFASIAS EN HABLA SEGUIDA. (Puntuar sobre si la longitud de la frase es de 4 palabras o más).	Presente en cada emisión.			1 – 2 casos por minuto de conversación.			Ausente.
6.	ENCONTRAR PALABRAS EN RELACION A LA FLUIDEZ DEL HABLA.	Habla fluida pero vacía.			Palabras informativas proporcionales a la fluidez.			Fácilmente obtiene palabras de contenido.
7.	REPETICIÓN DE ORACIONES. Puntuación percentil.	0-20	30	40	50	60	70-80	90-100
8.	COMPRENSIÓN AUDITIVA. Percentil medio de los 3 subtests estandar.	0-20	30	40	50	60	70-80	90-100

Volumen.	HIPOFONICO	NORMAL	FUERTE
Voz.	SUAVE	NORMAL	RONCA
Velocidad.	LENTA	NORMAL	RAPIDA

Índice de Competencia del Lenguaje.	

* *Basado en Goodglass, 2005.*

1

C.4 LISTADO DE PALABRAS OBJETIVO (DULCINEA)

Anexo V. LISTADO DE PALABRAS DEFINITIVO (Dificultad y Categoría)

	CONVERSACIÓN	
MONOSÍLABOS	BISÍLABOS	MÁS DE DOS
Sí / No Ya Más Mal Sin Con Qué / ¿Qué? Hoy	Gracias Adiós Hola Hora ¿Cuándo? ¿Cómo? (para hacer algo) Igual Quiero (pedir algo) Esta Esa ¡Basta! ¡Para! Menos	presente / pasado / futuro Problema Deseo... (pedir algo) Escúchame Necesitaría Mañana Ayuda
FRASES IMPERATIVAS	FRASES INTERROGATIVAS	FRASES DE CORTESÍA// CADENAS SOBREAPRENDIDAS
Vamos a pasear vamos a comer Vamos al cine Ven conmigo Dame un beso Quiero comer Quiero salir	¿Quién es? ¿Cómo estás? ¿Me entiendes? ¿Cuántos años tienes? ¿Qué hora es?	Buen día Buenas noches Por favor
FRASES DECLARATIVAS		
Yo quiero Yo tengo Necesito estar solo No te entiendo Voy al baño Voy a ducharme Voy a dormir Me ha pasado Te quiero Es muy importante		

	ESTADOS	
BISÍLABOS	MÁS DE DOS	FRASES DECLARATIVA
dolor feliz triste hambre Frío Calor Gota	pastilla mareo ansiedad	Me duele Me encuentro Me siento Me siento triste estoy contenta

	PERSONAS	
MONOSÍLABOS	BISÍLABOS	MÁS DE DOS
Yo Tú Él	mamá papá hijo hija abuelo abuela nieto nieta	Hermano Hermana Cuñado

	CUERPO	
MONOSÍLABOS	BISÍLABOS	MÁS DE DOS
pie	Ojo Mano Pierna	Cabeza Corazón

	COMIDA	
MONOSÍLABOS	BISÍLABOS	MÁS DE DOS
Pan Sal	Pera Carne Pollo Sopa Atún Limón Vino Agua Rico	Naranja Plátano Lechuga Ensalada Pescado Tomate Botella Cerveza Ensaladilla Lentejas Caliente

	VERBOS	
MONOSÍLABOS	BISÍLABOS	MÁS DE DOS
Dar Ven	Bañar Duchar Peinar Vestir Salir Comer Coger Leer	olvidar

	CASA	
MONOSÍLABOS	BISÍLABOS	MÁS DE DOS

gel	Plato Mesa Silla Vaso Copa Jarrón Sofá Tele Jabón Champú Peine Baño Cama	Cuchara Tenedor Cuchillo Escoba Toalla

	**SERIES**	
NÚMEROS	DÍAS SEMANA	MESES AÑO
Uno Dos Tres Cuatro Cinco Seis Siete Ocho Nueve Diez	Lunes Martes Miércoles Jueves Viernes Sábado Domingo	Enero Febrero Marzo Mayo Junio Julio Septiembre Octubre Noviembre Diciembre

	**LUGARES**	
	MÁS DE DOS	
	Gimnasio Piscina	

	**NATURALEZA**	
	MONOSÍLABAS	
	Flor Sol Pez	

	**ROPA**	
	BISÍLABAS	MÁS DE DOS
	Bolso Braga	Camisa Pantalón

		Calzoncillo

	OBJETOS	
	BISÍLABAS	
	Coche	

19 palabras que, a pesar de ser elegidas en la encuesta por los pacientes, finalmente no se incluyen en el listado de palabras a trabajar, por falta de escenas.

200 palabras en total contando esas 19

Por lo tanto, 181 palabras en total, de las que se disponen 831 escenas, existiendo aproximadamente entre 3-5 escenas por cada palabra.