Memoria Anual de Actividades de la Cátedra RTVE – UAH: Inteligencia Artificial y Accesibilidad en el Sector Audiovisual

Curso 2024-25

Contenido

1. Introducción

El 1 de agosto de 2024 se firmó el Convenio Específico de Colaboración entre la Corporación Radio y Televisión Española, S.A. y la Universidad de Alcalá (UAH) para la Creación de la Cátedra RTVE Inteligencia Artificial (IA) y Accesibilidad en el Sector Audiovisual. En la cláusula segunda de este convenio figuran las actividades a realizar, que son las siguientes:

- Análisis y evaluación de la calidad y de la sincronización actual de los subtitulados en RTVE: detección e identificación de la problemática existente. Para ello se procederá a la Recopilación de datos de RTVE (audios, subtítulos, etc.) que permita la construcción de distintos conjuntos de datos.
- Explorar la aplicación de tecnologías de transcripción para la comprobación del correcto funcionamiento de los subtítulos en una emisión determinada.
- **Realizar un estado del arte y estudio sobre las posibilidades de generación de lengua de signos mediante Inteligencia Artificial Generativa.**
- Evaluación y estudio de posibles soluciones, mediante inteligencia artificial, que permitan la traducción de subtitulados a otras lenguas distintas del castellano.
- **Promover la realización de trabajos académicos y científicos en materia de inteligencia artificial y accesibilidad: Trabajos Fin de Grado, Trabajos Fin de Máster, Artículos científicos, etc.**
- **Búsqueda de posibles convocatorias de proyectos de investigación a las que puedan concurrir conjuntamente los diferentes participantes de la Cátedra.**
- Todas aquellas otras que, acordadas entre las partes, contribuyan a la consecución del objetivo de la Cátedra.

En la reunión de lanzamiento, realizada el 20 de noviembre de 2024 en Torre España, se acordó comenzar a trabajar en la línea de generación de lengua de signos mediante Inteligencia Artificial Generativa. Adicionalmente, como se detallará más adelante, paralelamente se han llevado a cabo actuaciones en otras dos líneas de las indicadas en el convenio. En el párrafo anterior se han marcado en negrita las actividades del convenio abordadas durante este primer año de la Cátedra. La presente memoria tiene por objetivo recoger los detalles de todas las actuaciones que se han llevado a cabo en el seno de la Cátedra durante este primer año.

2. Actividades realizadas durante el curso 2024-25

Durante el primer año de vigencia de la Cátedra se han llevado a cabo numerosas actuaciones, cumpliendo con tres de las actividades que estaban previstas en el convenio, además de otras adicionales.

En los siguientes subapartados se explica con más detalle cada una de ellas.

2.1. Reuniones de la Comisión Mixta de Seguimiento

En este primer año se han realizado reuniones de seguimiento con una periodicidad aproximada de tres semanas, dando lugar a un total de **8 reuniones**, entre las cuales 2

de ellas se llevaron a cabo de manera presencial, y el resto de forma virtual. Las reuniones presenciales fueron el 20 de noviembre de 2024, en las instalaciones de RTVE en Torre España, para el lanzamiento de la Cátedra; y el 23 de junio de 2025, en la biblioteca CRAI de la Universidad de Alcalá.

Durante estas reuniones, se iba informando a RTVE de los avances científico-técnicos que iban teniendo lugar por parte de la UAH, y se ponían en común diferentes temas de interés para ambas instituciones y para la Cátedra. Se han realizado actas de cada una de estas reuniones.

2.2. Reuniones de las Cátedras de RTVE

RTVE tiene 11 Cátedras con 10 universidades españolas, entre las que se encuentra la UAH que, hasta la fecha, ha sido la última en adherirse a esta red. Los directores de estas Cátedras se reúnen cada cierto tiempo, junto con la Subdirección Centro de Innovación y Observación del Conocimiento de RTVE, para tratar temas en común y de cierto interés para las Cátedras.

En este curso académico, se han llevado a cabo **4 reuniones generales**, de las cuales 3 fueron virtuales y 1 fue presencial, el 8 de julio de 2025 en la Real Academia de Ingeniería.

2.3. Web de la Cátedra

Se ha creado una página web de la Cátedra (Fig. 1), donde se informa de las actividades que se llevan a cabo en el seno de la misma y de la composición del equipo. Además, también tiene como objetivo dar difusión a las publicaciones que se puedan derivar, y a las noticias que puedan surgir con respecto a la Cátedra.

Se puede acceder a la web a través de la dirección https://catedrartve.uah.es

Fig. 1. Página web de la Cátedra

2.4. Trabajos académicos y científicos

Tal y como se ha explicado anteriormente, otra de las actividades que figuran en el convenio es promover la realización de trabajos académicos y científicos en materia de inteligencia artificial y accesibilidad. Durante este año se ha realizado la tutorización y dirección, con éxito, de un Trabajo Fin de Grado (TFG) en el Grado en Ingeniería en Sistemas de Información de la Universidad de Alcalá, titulado *Estudio y Aplicación de Modelos de Inteligencia Artificial Generativa en la generación de Lengua de Signos Española*. Este trabajo ha permitido avanzar, a su vez, en los resultados obtenidos en la línea de generación de lengua de signos mediante IA. Fue presentado y defendido el 18 de junio de 2025, con una calificación de 10, y fue propuesto para Matrícula de Honor.

Además, para tratar de complementar los recursos de los que dispone la Cátedra, se ha solicitado (con el mismo estudiante que presentó su TFG, ya que adquirió una experiencia muy valiosa durante su desarrollo y obtuvo una muy buena calificación en su trabajo) una beca de iniciación a la investigación de la UAH, titulada *Generación de Lengua de Signos con Inteligencia Artificial Generativa.* Dicha solicitud, a fecha de elaboración de esta memoria, sigue en evaluación y se resolverá, previsiblemente, hacia finales de este año (2025).

2.5. Solicitudes de proyectos

Es interés de la Cátedra, como figura en el convenio, acudir a convocatorias de proyectos de investigación en las que se puedan proponer trabajos sobre las líneas de investigación existentes en la Cátedra.

Durante este año, se han solicitado 2 proyectos:

- *Using Generative Artificial Intelligence to Create an Interpreter for Sign Language (GEN-AI4SIGN),* presentado a la convocatoria Social Research 2025 de La Caixa.
- *Sistema avanzado de traducción automática de texto a lengua de signos española mediante inteligencia artificial generativa (SATALSE)*, presentado a la convocatoria 2024 de Proyectos de Generación de Conocimiento del Ministerio de Ciencia, Innovación y Universidades.

La solicitud de cada una de estas propuestas requiere la elaboración de una memoria, en ocasiones bastante extensa (unas 20 páginas), y suelen incluir un estado del arte, además de una descripción detallada de los objetivos, paquetes de trabajo, planificación, recursos, etc. Desafortunadamente, ambas propuestas no fueron financiadas, pero cabe destacar el esfuerzo que se ha puesto en realizar las solicitudes, ya que requiere de bastante tiempo y dedicación elaborar cada una de las propuestas.

2.6. Avances científico-técnicos

Generar avatares de lengua de signos mediante IA generativa es todo un reto. Hasta donde se ha podido investigar, no existe aún ningún estudio que haya conseguido realizar, con éxito, el proceso completo de convertir un texto (por ejemplo, a partir de subtítulos

de un vídeo) en un avatar. Este problema se vuelve aún más complejo y desafiante cuando se trata de la lengua española, donde los recursos son todavía más limitados que en otros idiomas como el inglés.

Para abordar este problema, se han ido estudiando los posibles enfoques diferentes (Fig. 2.) y se ha ido trabajando en cada uno de ellos.

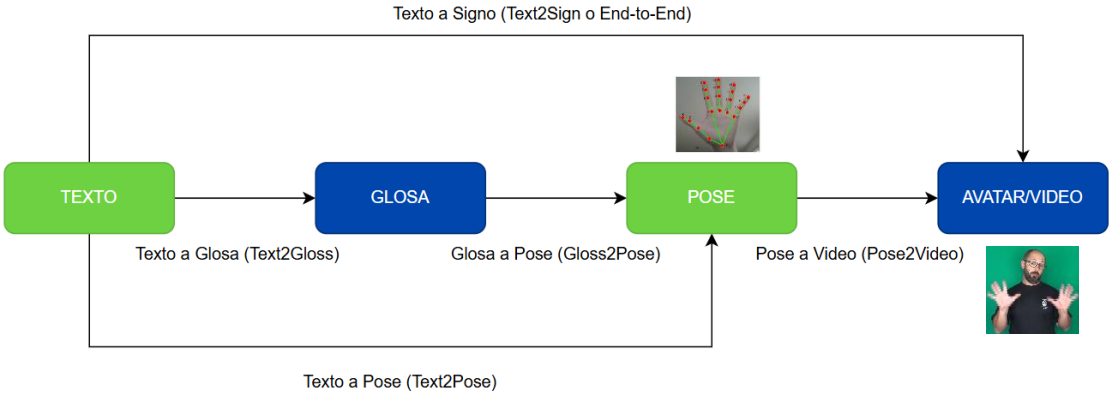

Fig. 2. Diferentes enfoques para resolver el problema

El trabajo se ha dividido en dos partes: (1) convertir de texto a pose y (2) convertir de pose a vídeo. En un principio se ha ido trabajando más en la línea de obtener una pose, pero paralelamente se ha ido avanzando también en el estudio de conversión de poses a avatares, aunque esta parte está menos madura.

El primero de los enfoques sería el más tradicional, que consiste en pasar de texto a glosa (Text2Gloss), luego de glosa a pose (Gloss2Pose) y, por último, de pose a avatar (Pose2Video). Este enfoque se explica más en detalle en el apartado 2.6.2. También existiría la posibilidad de hacer una conversión directa del texto a un avatar (Text2Sign), sin pasar por pasos intermedios. Este es un enfoque bastante innovador y, por tanto, bastante retador y con alta probabilidad de fracaso. Se explica en el apartado 2.6.3. Por último, existe un enfoque mixto que consistiría en convertir de texto a pose (Text2Pose) y, posteriormente, de pose a avatar (Pose2Video). Este enfoque se explica en el apartado 2.6.1.

2.6.1. Enfoque Texto a Pose (Text2Pose)

Dentro de este enfoque, se han probado diferentes opciones para tratar de conseguir los mejores resultados posibles. No obstante, con ninguno de ellos se ha llegado a conseguir una pose que corresponda al signo indicado en el texto original.

2.6.1.1 Modelo Decoder ajustado para predecir poses

Explicación corta:

Modelo Transformers con un Decoder pre-entrenado Qwen para predecir la pose de un frame de forma autorregresiva.

Explicación más detallada:

Este modelo ha sido la principal aplicación que se ha implementado, similar a modelos del estado del arte como Progressive Transformers. Se basa en aplicar la arquitectura Transformers con un modelo pre-entrenado para predecir los frames del vídeo de forma autorregresiva.

El primer paso es procesar los datos, tanto el texto como los landmarks (puntos clave de la pose). Para ello, se aplica un encoder de texto DeBERTa-v3 para generar los vectores de embeddings de los tokens del texto y embeddings especializados para los landmarks. Como los embeddings del texto tienen diferentes dimensiones que la entrada del Decoder, para ello también se ha aplicado una red Linear como adaptador para ajustar esta dimensionalidad.

En los embeddings de landmarks se aplicaron al principio, además de embeddings posicionales aprendidos para que la pose contuviera más información y no se produjera mucha pérdida de información. Sin embargo, estos embeddings de posición generaban vibración y no resultaban beneficiosos en comparación con no utilizarlos, posiblemente debido a que Qwen ya incorpora información posicional implícita a través de RoPE (Rotary Position Embedding), un método que codifica información posicional rotando las representaciones en el espacio de embeddings, haciendo redundantes los embeddings posicionales adicionales. Por otro lado, para separar las partes del cuerpo en los embeddings se aplicaron varias capas Linear, una para cada parte del cuerpo (cabeza, cuerpo, mano izquierda, mano derecha).

El Decoder Qwen recibe los embeddings de texto y los embeddings de landmarks separados por un token de inicio de secuencia para que el modelo tenga la información sobre cuándo comienzan los landmarks. Además, se añade un token de final de secuencia (EOS) que se implementa como una dimensión adicional en los puntos de los landmarks donde un valor de 1 indica que ese frame es el final de secuencia y 0 que continúa. Durante el entrenamiento, esto genera dos pérdidas: una MSE (Mean Squared Error) que mide la diferencia cuadrática entre las coordenadas predichas y reales de los landmarks, y una BCE (Binary Cross Entropy), que evalúa la precisión de la clasificación binaria del token de EOS. El decoder predice los frames de forma autorregresiva utilizando *teacher forcing* durante el entrenamiento. El *teacher forcing* es una técnica donde el modelo recibe la secuencia ground truth real y aprende a predecir el siguiente frame basándose en los frames reales anteriores, en lugar de usar sus propias predicciones. Durante la inferencia, el modelo predice un frame y luego vuelve a realizar el proceso con la entrada de texto y el frame predicho para generar el siguiente, y así procede hasta que finaliza de generar el vídeo.

La justificación de este enfoque se basa en la naturaleza intrínsecamente secuencial y autorregresiva del lenguaje de signos. Al igual que el lenguaje escrito o hablado, el lenguaje de signos presenta una estructura temporal donde cada gesto o pose depende del contexto anterior, lo que sugiere que las técnicas de modelado de lenguaje natural deberían ser efectivas para la generación de secuencias de poses.

Problemas identificados:

A pesar de que la teoría es atractiva, se han identificado varios problemas durante el desarrollo y entrenamiento:

- **Regresión a la media (Regression to the mean):** El modelo tiende a generar poses que convergen hacia posiciones promedio o neutras, perdiendo la variabilidad y expresividad natural del lenguaje de signos. Técnicamente, esto ocurre porque la función MSE penaliza cuadráticamente las desviaciones, incentivando al modelo a predecir el valor esperado de la distribución de poses. En la práctica, esto se traduce en manos que permanecen semi-cerradas, brazos que no alcanzan extensiones completas y expresiones faciales neutras.
- **Desarticulación esquelética**: El modelo puede generar poses donde las articulaciones y conexiones anatómicas del esqueleto humano no se mantienen de forma coherente, especialmente en las manos. Técnicamente, esto ocurre porque el modelo trata cada coordenada como una regresión independiente sin restricciones cinemáticas explícitas. Las manos sufren particularmente porque tienen 21 puntos (en todos los formatos) en un espacio muy reducido, y pequeños errores de predicción pueden resultar en dedos que atraviesan la palma o muñecas rotadas 360 grados. Este problema se acentúa hacia el final del vídeo, cuando el error acumulado de predicciones autorregresivas hace que el esqueleto parezca más un "churro" que una persona.
- **Dificultad para capturar movimientos dinámicos**: Los gestos rápidos o con alta variabilidad espacial pueden ser suavizados excesivamente, resultando en movimientos menos expresivos.
- **Balanceo de pérdidas**: La combinación de MSE y BCE requiere un ajuste cuidadoso de pesos para evitar que una pérdida domine sobre la otra, afectando tanto la calidad de las coordenadas como la detección del final de secuencia. Técnicamente, se está optimizando simultáneamente una tarea de regresión continua (x coordenadas) y una clasificación binaria (token EOS), con escalas de pérdida muy diferentes. Además, la pérdida BCE presenta un desbalance de clases muy alto (el 99% de los ejemplos son "no es el final" y solo el 1% son "sí es el final"). El modelo aprende rápidamente que decir siempre "no" le da un 99% de accuracy, pero en la práctica nunca termina de generar frames y acaba con vídeos infinitos de un esqueleto moviéndose a cámara lenta haciendo el mismo gesto (o al menos lo intenta) por siempre. Para tratar de minimizar esto, al final limitamos manualmente el número de frames que genera.
- **Pérdida de alineación texto-pose:** El problema más fundamental es que los signos generados frecuentemente no corresponden al texto de entrada. El modelo puede generar secuencias de poses que parecen movimientos de lengua de signos, pero no representan la palabra o frase que debería. Esto sugiere que el entrenamiento no logró crear una asociación robusta entre los embeddings de texto y las secuencias de poses correspondientes. La función de pérdida optimiza la verosimilitud de las poses, pero no garantiza que estas poses tengan el significado semántico correcto. El modelo genera movimientos que parecen algo,

pero no se parece a lo que debería. Es posible, incluso, que ni siquiera sea un signo válido.

Resultados:

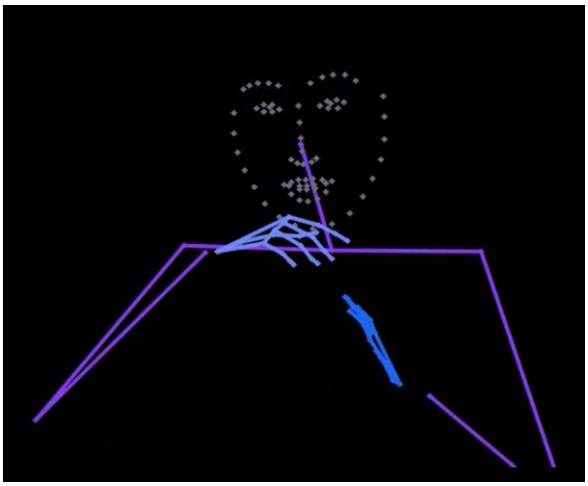

Fig. 3. Captura de vídeo generado con el modelo Decoder ajustado para predecir poses

2.6.1.2 Predecir frames

Explicación corta:

Se ha cambiado la cabeza del Decoder de Qwen para predecir varios frames (2, 3, 5, 10, 25...) de forma autorregresiva en lugar de un solo frame, con el objetivo de mejorar los resultados obtenidos.

Explicación más detallada:

Este enfoque surge como una respuesta a los problemas identificados en el método puramente autorregresivo, especialmente la regresión a la media y la desarticulación esquelética que se acentúa con la generación secuencial de frames individuales.

La modificación principal consiste en cambiar la dimensión de salida de la cabeza final del modelo para generar los siguientes n frames simultáneamente en lugar de generar un frame individual por paso. Técnicamente, esto implica modificar la última capa lineal del modelo para que produzca un tensor de salida con dimensión [batch, seq_len, features * n], donde features incluye las coordenadas de todos los puntos corporales más el token EOS. Esta configuración es ajustable y puede tomar valores como 2, 3, 5, 10 o 25 frames por paso.

Durante el entrenamiento, debido a la naturaleza de los Transformers, que procesan secuencias completas en paralelo mediante atención causal, cuando se le pasa un clip completo de 50 frames al modelo y se configuran 5 frames por paso, el modelo genera una predicción para cada posición de entrada. Específicamente, para cada posición temporal t, el modelo predice los siguientes 5 frames [t+1, t+2, t+3, t+4, t+5], resultando en 50 × 5 = 250 frames generados en total. La máscara de atención causal garantiza que cada predicción solo pueda acceder a los tokens anteriores, manteniendo la propiedad autorregresiva. La función de pérdida se aplica a todos estos frames generados, donde

cada frame predicho se compara con su correspondiente frame ground truth desplazado temporalmente. Esto significa que todos los 250 frames contribuyen al gradiente en cada paso de retropropagación, permitiendo un aprendizaje más eficiente al aprovechar el paralelismo inherente de la arquitectura Transformer.

Durante la inferencia, el proceso es diferente y verdaderamente autorregresivo: se comienza solo con el texto como entrada y un token BOS (Beginning of Sequence) para los landmarks. El modelo genera los primeros 5 frames, estos se procesan a través de los embeddings de landmarks (separados por región corporal) y se concatenan al contexto existente. Luego, el modelo genera los siguientes 5 frames basándose en el texto y todos los frames anteriores, repitiendo este proceso hasta detectar el token EOS o alcanzar la longitud máxima. Si se quieren generar 50 frames, se necesitarían 10 pasos autorregresivos (50 ÷ 5) en lugar de los 50 pasos que requeriría el enfoque de frame único, reduciendo significativamente la latencia de generación.

Problemas identificados:

Este enfoque también presenta nuevos desafíos:

- Discontinuidades temporales: Al generar bloques de n frames de manera simultánea, se pueden producir cortes abruptos cada n frames durante la inferencia. Técnicamente, el último frame del bloque k y el primer frame del bloque k+1 no tienen ninguna restricción de continuidad durante el entrenamiento, ya que fueron generados en forward passes diferentes. Esto resulta en saltos visuales cada n frames donde las manos pueden teletransportarse o los codos cambian de ángulo.
- Mayor complejidad de predicción: El modelo debe aprender a predecir múltiples frames futuros simultáneamente, lo que aumenta exponencialmente la dificultad de la tarea. Técnicamente, en lugar de aprender P(frame_t+1 | contexto), ahora debe aprender P(frame_t+1, frame_t+2, ..., frame_t+n | contexto), un espacio de salida mucho más complejo. El resultado es que el modelo puede generar secuencias coherentes internamente dentro de cada bloque, pero pierde precisión en los detalles finos de cada frame individual, especialmente en los frames más lejanos del bloque.

Resultados:

Los resultados no son significativos, ya que no mejoran al modelo Decoder de Qwen cuando genera solo un frame por paso.

2.6.1.3 Conjunto finito de poses

Explicación corta:

Se ha ajustado el modelo entrenado del Decoder con Qwen con un conjunto de datos finito para que sea capaz de generar ciertas palabras de un vocabulario reducido, en lugar de diferentes palabras de un vocabulario mucho más amplio.

Explicación más detallada:

Este enfoque representa un intento de simplificar el problema de generación de poses mediante la especialización del modelo en un subconjunto específico de signos. En lugar de intentar que el modelo aprenda a generar cualquier signo del lenguaje de signos completo, se realiza un fine-tuning del modelo Qwen ya entrenado (el del primer enfoque que genera un frame por paso) sobre un conjunto de datos mucho más reducido y específico.

Para el conjunto de datos reducido se han utilizado datos de poses en las que se signan palabras de animales. Como los resultados tenían algunos avances en palabras como "perro", pero tenían limitaciones como la duración o que no realizaba bien signos como "zorro", se creó, a partir del conjunto facilitado por RTVE, un conjunto de datos reducido de una palabra y sus variantes como, por ejemplo, "familia", con variantes como "familiar".

La justificación de este enfoque se basa en un intento de especialización. Al reducir drásticamente el espacio de hipótesis de miles de signos posibles a solo un conjunto de palabras específicas, se esperaba que el modelo pudiera memorizar mejor las secuencias específicas de cada signo, reducir la ambigüedad en la generación al tener menos opciones posibles, mejorar la consistencia temporal al entrenar con ejemplos más homogéneos, y facilitar el aprendizaje, al convertirlo en un problema prácticamente de recuperación en lugar de generación creativa. Básicamente, era un intento de lograr un mapeo palabra-signo generado por el modelo.

Problemas identificados:

Este enfoque tampoco está exento de problemas:

- Mejora mínima sobre el modelo base: A pesar de la especialización, las mejoras fueron marginales. El modelo seguía presentando los mismos problemas de regresión a la media y desarticulación esquelética.
- Persistencia de problemas fundamentales: A pesar de la especialización en un vocabulario reducido, el problema de alineación texto-pose se mantuvo. Los signos generados seguían sin corresponder al texto de entrada, lo que puede significar que la desconexión semántica es un problema arquitectural más profundo que no se resuelve simplemente reduciendo el espacio de salida.

Resultados:

En los vídeos generados para los distintos conjuntos finitos se encuentran resultados positivos, como que en alguna parte del vídeo se realizaban movimientos similares al signo objetivo. Sin embargo, en el resto del vídeo, las poses no tenían ninguna relación con el objetivo.

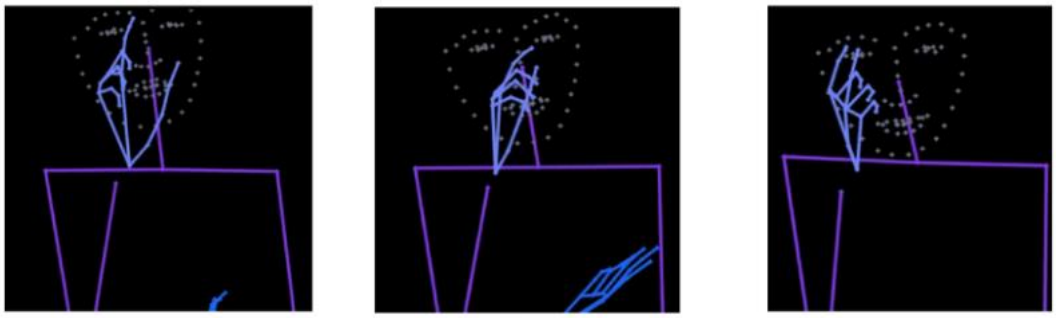

Fig. 4. Ejemplo de fotogramas del vídeo generado para la palabra "perro"

2.6.1.4 Modelo Transformers de Bajo Nivel

Explicación corta:

Este modelo consiste en un modelo Transformer de bajo nivel, donde no se ajusta un modelo pre-entrenado, sino que se cambian ciertos módulos de la arquitectura Transformer, como el módulo de Cross Attention, para que el modelo tenga más en cuenta los tokens de texto durante todo el entrenamiento.

Explicación más detallada:

Es un modelo que aplica una arquitectura Transformer sin un modelo pre-entrenado como Qwen en el Decoder. Un problema del Modelo Decoder ajustado para predecir poses (explicado en la sección 2.6.1.1) es la diferencia de cantidades entre tokens de texto y los puntos clave de la pose durante el entrenamiento, siendo mucho mayor el número de puntos. Esta diferencia puede implicar que el modelo no utilice los tokens de texto cuando se han llevado a cabo muchos pasos en el entrenamiento. Por ello, en lugar de utilizar un Encoder de texto, se aplica en el Decoder un nuevo módulo de atención cruzada que asegura el uso de los tokens de texto durante todo el entrenamiento.

Resultados:

Los resultados de este modelo mejoraron la fluidez del vídeo generado respecto al Modelo Decoder ajustado para predecir poses (explicado en el apartado 2.6.1.1). Sin embargo, en cuanto a la traducción, este modelo no logra mejorar la obtención de la secuencia de poses adecuadas para un signo.

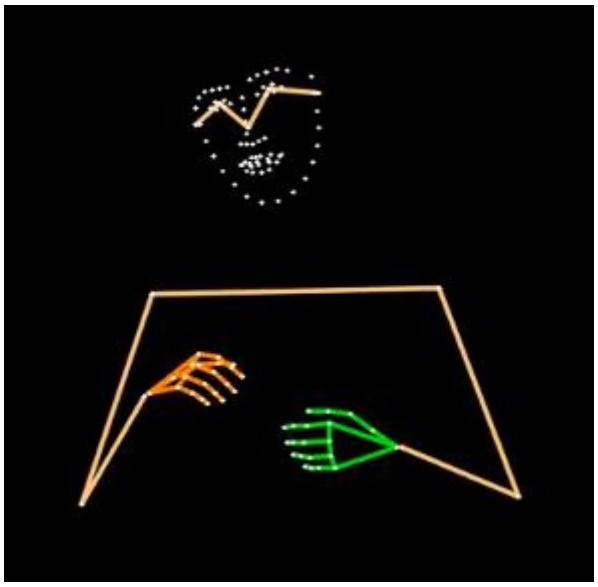

Fig. 5. Captura del vídeo generado por el modelo Transformer de Bajo Nivel

2.6.1.5 Diferentes pérdidas

Explicación corta:

Durante los entrenamientos, se han empleado diferentes pérdidas para que los modelos fueran capaces de aprender, no solo la pérdida de las diferencias entre los puntos de los landmarks, sino también pérdidas de diferencia temporal, de aceleración, para predecir el final o para conocer con un contador el frame generado.

Explicación más detallada:

En todos los modelos que se han desarrollado se han utilizado las siguientes pérdidas para que el modelo pudiera ajustar sus pesos y mejorar los resultados obtenidos:

- Pérdida Landmarks: entropía cruzada entre los puntos generados y los puntos de la pose correspondiente.
- Pérdida Temporal: primera derivada de la pérdida de los landmarks. Sirve para tratar de mejorar el movimiento de la pose.
- Pérdida Acceleration: segunda derivada de la pérdida de los landmarks. Se utiliza para mejorar la suavidad de los movimientos en los vídeos generados.
- Pérdida Fin del vídeo: entropía binaria entre el último token y el token EOS para calcular si se debe continuar generando un frame o si el vídeo ya ha finalizado.
- CounterEmbedding: basado en Progressive Transformers, se asigna un último token a cada frame y se normaliza de forma que el último frame tenga un "1" como último token. De esta manera, se podrá predecir cuándo acaba el vídeo.

2.6.2. Enfoque Texto a Glosas (+ Glosas a Pose)

Este enfoque consiste en la utilización de glosas como elemento intermedio que facilite la generación de poses. Para este enfoque se realizan aproximaciones tanto con modelos ajustados para generar glosas como con LLMs que reciben una frase y generan su frase con glosas.

2.6.2.1 Glosas

Explicación corta:

Se ha utilizado un Decoder pre-entrenado para predecir las glosas a partir de un texto, utilizando un conjunto de datos de texto-frases extraído de GlossLM.

Explicación más detallada:

A partir del conjunto de datos de GlossLM, se ha obtenido un conjunto de datos con frases en lenguaje español escrito y las glosas correspondientes. Este conjunto total tiene un tamaño de 22.369 pares de texto y glosas, que han servido para ajustar a un modelo pre-entrenado. El modelo pre-entrenado es un modelo Decoder que tiene como objetivo generar glosas a partir de una frase.

Para el entrenamiento del sistema, se ha utilizado un modelo pre-entrenado de tipo Decoder, específicamente diseñado para tareas de generación de secuencias. El objetivo principal de este modelo es producir automáticamente glosas a partir de una frase en español, lo que puede ser de gran utilidad en los modelos de generación de pose para reducir información redundante, como palabras que no aporten información en Lengua de Signos Española (LSE) y no tengan traducción.

Problemas identificados:

El proceso de entrenamiento incluyó técnicas de fine-tuning para adaptar el modelo a la estructura particular de las glosas. Sin embargo, los resultados generados por el modelo no fueron lo suficientemente prometedores como para aplicar este modelo en la traducción de texto a glosa del dataset utilizado con las frases proporcionadas por RTVE, tampoco sirviendo para el entrenamiento de los modelos anteriores.

Resultados:

Los resultados de este modelo presentaban ciertas inconsistencias, como que seguía generando frases en inglés o que proporcionaba resultados cortos e inexactos para algunas frases. Por ejemplo, para la frase "Ha faltado algo de tiempo", el resultado era "decady".

2.6.2.2 Modelo con LLM

Explicación corta:

Modelo basado en la metodología explicada en el paper "*Towards AI-driven Sign Language Generation with Non-manual Markers*", https://arxiv.org/abs/2502.05661), personalizado para Lengua de Signos Española. Para ello, primero se probó a generar glosas con consultas con GPT-4o, como en el paper original. Sin embargo, debido a las

limitaciones que presenta el uso de este modelo, en este caso se ha utilizado Qwen para automatizar la generación del vídeo.

Explicación más detallada:

Este modelo consiste en utilizar un Gran Modelo de Lenguaje (LLM) como generador de glosas. Las glosas que genera el modelo están limitadas a un vocabulario, puesto que para generar glosas se realiza una serie de consultas previas donde se indica qué son las glosas, se proporcionan frases con ejemplos y un vocabulario que manejar.

En un primer momento, se ha utilizado ChatGPT para probar la generación con el modelo de GPT-4o, tal y como se hace para el modelo en inglés en el trabajo publicado mencionado anteriormente. Los resultados ofrecidos son bastante positivos y no suele generar glosas fuera del vocabulario, lo que es destacable.

Sin embargo, el modelo de GPT-4o presenta dos grandes limitaciones. La primera de ellas es su acceso ilimitado de consultas mediante pago. Esto impide la utilización de una API para automatizar la generación del vídeo de manera gratuita, por lo que habría que realizar el proceso de forma manual, generando las glosas en ChatGPT y teniendo que pegarlas posteriormente en una interfaz para generar el vídeo, lo cual es inviable. Además, no tiene un buen manejo para la generación de frases largas que contengan más de cuatro palabras. Debido a estas limitaciones se han explorado otros modelos como Gemma y Qwen para automatizar este proceso, con un prototipo sencillo, haciendo uso de Streamlit para la interfaz y de vLLM para la interacción con los modelos que además permiten manejar la oración de entrada para ofrecer ayuda en las consultas al modelo para que su generación sea precisa. Tras explorar estos modelos, se ha seleccionado gemma/gemma-2-2b para suplir la función de GPT-4o.

Este nuevo prototipo resulta menos efectivo en la generación que el modelo de GPT-4o, sobre todo para palabras desconocidas del vocabulario. Por ello, se realiza una consulta diferente donde se indica si la palabra pertenece al vocabulario, para que sea capaz de generarla; o si la palabra no pertenece al vocabulario, para que realice una paráfrasis con el vocabulario ofrecido y genere glosas válidas para crear el vídeo. En este caso, se divide la frase en una lista de palabras para enviarlas al modelo una a una y mejorar la obtención de glosas respecto a enviar la frase en lotes de palabras.

Problemas identificados:

Para expresiones como "cada año", ni GPT-4o ni Gemma son capaces de generar las glosas de forma correcta. Por ello, con el objetivo de facilitar la generación, se traducen directamente expresiones de dos palabras o más del diccionario de glosas.

Por último, se han valorado casos en los que, tras 1 o 2 consultas al modelo, este no sea capaz de generar la glosa con el vocabulario ofrecido. Para este tipo de casos límite, se comprueba que la salida de glosas no es válida y se genera una glosa deletreando en mayúsculas la palabra y concatenando con guiones, para imitar un proceso de dactilografía. Este proceso es útil también para nombres propios como "Manuel", cuya glosa se genera como "M-A-N-U-E-L". La Fig. 6 muestra cómo se ve la interfaz cuando se han generado las glosas y el vídeo correspondiente para un texto.

Resultados:

En los vídeos generados por este modelo, se consiguen los mejores resultados de traducción con una obtención de glosas para palabras del conjunto de datos y con la obtención de varias glosas que actúan como sinónimos cuando la palabra se encuentre fuera del vocabulario.

Fig. 6. Ejemplo de la interfaz del Modelo de Traducción con LLM

2.6.3. Enfoque Texto a Pose con difusión

El objetivo de este enfoque es utilizar mecanismos de difusión, como la eliminación de la autorregresión, para la obtención de la pose a partir del texto. Para realizar este enfoque es importante conocer el número de frames del vídeo; por ello, se ha avanzado en la creación de un modelo para predecir este número.

2.6.3.1 Predecir el número de frames de una palabra (modelos NAT)

Explicación corta:

Tratar de predecir el número de frames que son necesarios en un vídeo para signar una palabra y utilizar después modelos No AuTorregresivos (NAT) para generar la secuencia de poses.

Explicación más detallada:

Para desarrollar modelos No Autorregresivos es necesario conocer el número de frames que son necesarios predecir para generar el vídeo, es decir, si queremos generar los signos para un texto de "Hola, ¿cómo estás?", primero tenemos que conocer cuántos frames ocupa dicho signo y luego predecir directamente la pose de todos estos frames.

Este modelo se basa en un Encoder de BERT, y está entrenado con pares de frases de texto y el número de frames. Además, el número de frames de cada texto sufrió un procesamiento para disminuir y ajustar los datos para poder ser entrenado por el modelo. En primer lugar, se probó a dividir las coordenadas en una décima parte para el entrenamiento y deshacer esta operación en el modo de inferencia, multiplicando por la décima parte el número predicho por el modelo. También, se aplicó una normalización de los valores, tanto a [0,1] como a [-1, 1]. Sin embargo, con las predicciones para frases cortas de pocos frames no se obtenían buenos resultados.

Resultados:

Los resultados, en un primer momento, no son los mejores y suele generar el número de frames en un rango similar para varias frases. Por tanto, primero se está tratando de realizar la generación con difusión de la pose con el número real de frames, para ver si el enfoque es posible, y después la idea es centrarse en este modelo.

2.6.3.2 Eliminación de la autorregresión

Explicación corta:

Se suprimió el mecanismo autorregresivo para generar todos los fotogramas del clip en un solo paso, pero al no existir dependencia explícita entre cada frame, el modelo no logra representar correctamente el gesto completo y las poses presentan un temblor más pronunciado.

Explicación más detallada:

En la versión autorregresiva, cada nuevo fotograma se basa en el anterior, lo que garantiza cierta coherencia temporal y suavidad en la transición de movimientos. Al eliminar este componente, se construyen todos los fotogramas simultáneamente sin referenciar las predicciones anteriores, perdiendo la correlación natural entre posiciones consecutivas. Esta falta de contexto provoca que el modelo no siga la secuencia gestual esperada, ya que no "sabe" cómo evolucionó la pose en el paso anterior. Además, la ausencia de dependencia frame a frame aumenta el ruido en las coordenadas predichas: pequeños errores se amplifican y cada fotograma carece de información continua, generando movimientos inconsistentes y temblores más notorios en las manos y el cuerpo.

Problemas identificados:

– Se observan vibraciones más intensas en las articulaciones, especialmente en manos y hombros, dado que no existe un suavizado natural entre frames consecutivos.
– Aunque el enfoque no autorregresivo es más eficiente computacionalmente, sacrifica la coherencia temporal, resultando en animaciones menos estables y realistas.

Resultados:

El modelo no autorregresivo permitió generar los fotogramas de forma más rápida, marcando mejoras en los tiempos de inferencia. Sin embargo, esta ganancia en eficiencia se presenta junto con una clara pérdida de coherencia temporal, generando un mayor movimiento de los landmarks, pero también mayores temblores y gestos irregulares. A

diferencia del enfoque autorregresivo, donde cada frame hereda información del anterior, se provoca una secuencia menos fluida y con mayor variabilidad entre fotogramas. En consecuencia, aunque el método mejora el rendimiento y la velocidad de generación, compromete la estabilidad visual y la naturalidad del movimiento.

Fig. 7. Ejemplo de frames generados sin autorregresión

2.6.4. Técnicas utilizadas en ambos enfoques

Dentro de este apartado se explican algunas técnicas utilizadas en uno o varios enfoques descritos. La mayoría de estas técnicas se centran en la mejora de la pose con la que se muestran los signos a realizar.

2.6.4.1 Diferentes herramientas de obtención de pose

Explicación corta:

Para los modelos de generación de signos, se generan poses que representan la figura humana, similares a un esqueleto. Para la extracción de estas poses para el entrenamiento de estos modelos se han empleado diferentes herramientas, como OpenPose, MediaPipe, COCO y DWPose.

Explicación más detallada:

Se han utilizado herramientas de extracción de puntos en 2D para obtener ficheros JSON con los puntos de las poses para el entrenamiento, en concreto:

- Media Pipe[1]: es un modelo de Google para extraer puntos de la pose humana. Genera 21 puntos para cada mano, 33 puntos para la pose y 468 puntos para la cara, sumando un total de 543 puntos generados. De las diferentes herramientas probadas, esta es la que más puntos corporales extrae, lo que aumentaba considerablemente la memoria utilizada para el dataset, así como durante el entrenamiento.
- OpenPose[2]: tiene 21 puntos para cada mano, 70 puntos para la pose y 25 para la cara (mucho menos que MediaPipe). Sin embargo, su procesamiento de tiempo es mayor, tardando 4,5 segundos en generar la pose por cada segundo de vídeo, lo que aumenta los requerimientos computacionales para grandes conjuntos de

[1] https://ai.google.dev/edge/mediapipe/solutions/guide?hl=es-419
[2] https://github.com/CMU-Perceptual-Computing-Lab/openpose

datos. No obstante, el hecho de tener menos puntos en la cara disminuye el ruido para tratar de generar los puntos de las manos y la pose de forma más efectiva.

- COCO[3]: Tiene un número similar de puntos a OpenPose, salvo en la cara, que extrae 29 puntos, pero su tiempo de procesamiento es menor. Además, tiene la posibilidad de valorar los puntos en 3D.
- DWPose[4]: Es un modelo derivado de MMPose que extrae un total de 133 puntos, cantidad muy similar a Openpose. Reduce considerablemente el tiempo de obtención y la memoria utilizada frente a MediaPipe, manteniendo un nivel de detalle suficiente para capturar la expresividad facial y precisión en las manos, algo esencial en el modelado de signos.

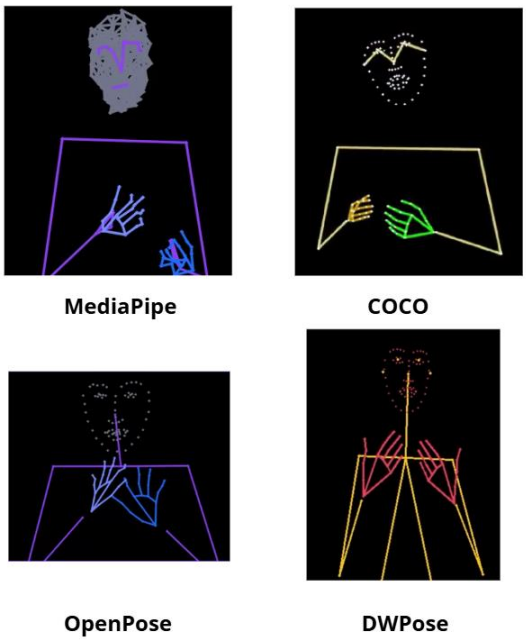

Fig. 8. Representación de una pose mediante diferentes herramientas

2.6.4.2 Normalización

Explicación corta:

Se implementaron múltiples estrategias de normalización para escalar las coordenadas de los puntos de la pose, asegurando coherencia espacial y mejorando la estabilidad durante el entrenamiento. Las opciones probadas incluyen la normalización frame a frame o global al vídeo, y distintos rangos como [0,1] o [-1,1].

Explicación más detallada:

Con el objetivo de reducir la variabilidad espacial entre las poses de dos personas se incorporaron técnicas de normalización sobre las coordenadas 2D de los landmarks extraídos. En concreto, se desarrollaron dos enfoques principales:

[3] https://github.com/jin-s13/COCO-WholeBody
[4] https://github.com/IDEA-Research/DWPose

- Normalización por frame: Cada frame se normaliza individualmente tomando como centro la media entre los hombros (teniendo en cuenta los índices específicos para los mismos, tanto en OpenPose como en MediaPipe), y escalando en función de la distancia entre ellos. Este método conserva la estructura relativa de cada frame, pero puede introducir pequeñas discontinuidades temporales al no considerar la secuencia completa.
- Normalización global al vídeo: Se calcula un centro y una distancia media de hombros a lo largo de todos los frames del vídeo. Posteriormente, todos los frames se normalizan con respecto a estos valores globales. Esta técnica preserva mejor la continuidad temporal y ofrece una referencia espacial común para todo el clip.

Ambos métodos transforman las coordenadas a un espacio centrado en el torso, utilizando escalas relativas a la anchura de los hombros para garantizar invariancia al tamaño de la persona o a la resolución del vídeo. Se probaron también distintas formas de normalización estándar (escalar a [0,1] o a [-1,1]), aunque finalmente se optó por una transformación centrada y escalada respecto a los hombros.

2.6.4.3 Filtrar vídeos para mejorar la pose

Explicación corta:

Las poses utilizadas para los entrenamientos a veces no tienen la calidad suficiente, por lo que se han aplicado filtros a los puntos para mejorar el flujo en las poses.

Explicación más detallada:

Dado que en algunos casos el uso de herramientas para la extracción de puntos genera secuencias de puntos poco entrelazados y con baja calidad, se ha realizado un filtrado de los puntos para eliminar el ruido que pudiera existir y conseguir una secuencia de puntos más coherente y cohesionada durante todo el tiempo del vídeo y con los puntos de las manos más visibles. Uno de los filtros utilizados ha sido el filtro de Savitzky–Golay, aplicado en otras áreas para filtrar ruido, como en los registros de electrocardiogramas (ECG). Se trata de un filtro basado en una regresión polinomial local para determinar nuevos puntos y añadir frames adicionales para mejorar las transiciones de frames o para obtener la posición de puntos que quedaran ocultos.

El objetivo de este filtrado es el de mejorar el dataset del entrenamiento de los modelos para que estos se entrenen con los mejores datos posibles. Además, estas técnicas de filtrado se pueden aplicar a los vídeos generados disminuyendo la inconsistencia temporal de los frames generados por el modelo.

Resultados:

En los resultados de la aplicación de este filtro se observa una mejor fluidez del vídeo respecto con el original, aunque la diferencia entre ambos vídeos no es significativa, pero el filtro mejora ciertas transiciones.

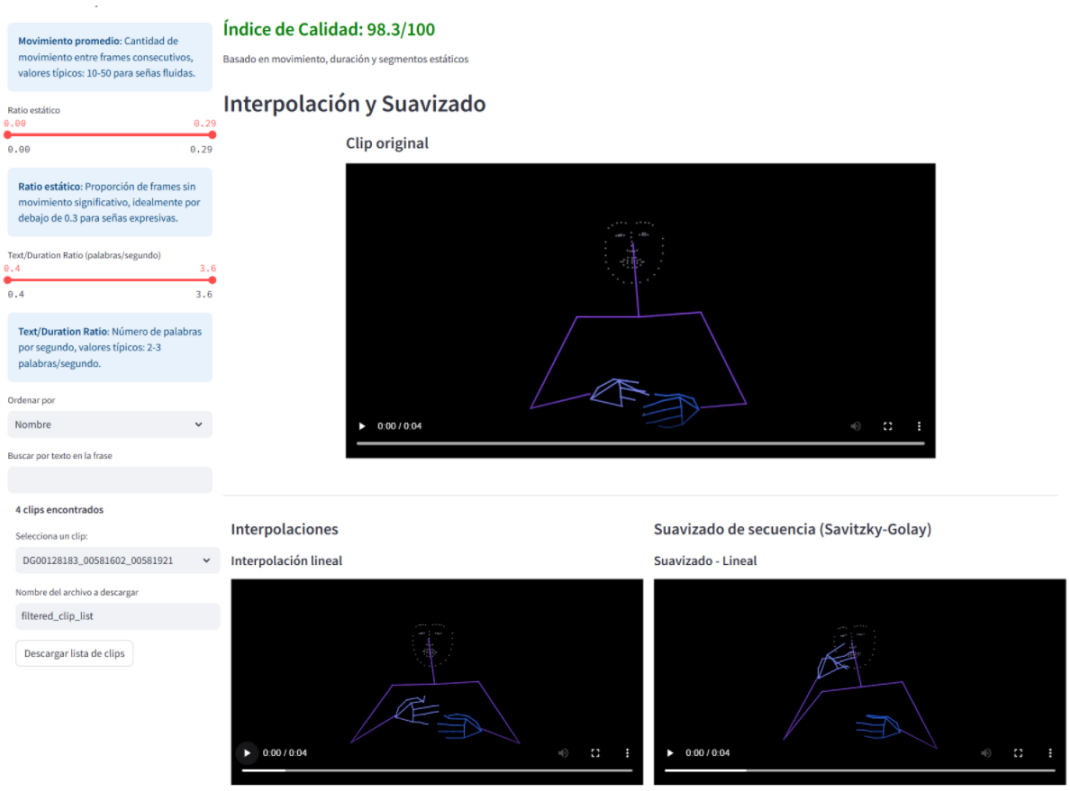

Fig. 9. Ejemplo de interfaz para visualización con interpolación.

2.6.4.4 G2P-DDM e interpolación de puntos

Explicación corta:

Se exploró el uso del modelo G2P-DDM para convertir puntos en 2D a 3D, así como técnicas de interpolación para mejorar la continuidad temporal en los puntos clave, pudiendo recuperar información faltante por baja calidad en los vídeos.

Explicación más detallada:

Con el objetivo de mejorar la calidad de las poses generadas, se evaluó la posibilidad de aplicar el modelo G2P-DDM, una arquitectura originalmente pensada para tareas de generación de gestos a partir de texto (Gesture-to-Pose) que incluye un módulo para predecir puntos en 3D a partir de una representación 2D. Sin embargo, al aplicar este modelo sobre nuestras secuencias 2D extraídas mediante OpenPose o MediaPipe, los resultados obtenidos en 3D fueron de baja calidad, generando poses inconsistentes.

Pese a ello, se decidió rescatar parte de la idea del modelo, en concreto, la interpolación de puntos, y aplicar técnicas similares sobre los puntos 2D extraídos. La idea era suavizar la trayectoria de los puntos y mejorar la fluidez entre frames, así como predecir posible información faltante u obtenida con un valor de confianza muy bajo. Sin embargo, debido a que las herramientas ya utilizadas (como OpenPose y MediaPipe) ofrecían una buena precisión en la mayoría de los casos, las mejoras conseguidas con la interpolación fueron mínimas y no compensaban el coste computacional añadido.

Resultados:

Se evaluó el impacto de la interpolación sobre secuencias de baja calidad visual utilizando la base de datos PHOENIX-2014-T, dado su tamaño más reducido y manejable, así como la presencia de numerosos frames con detección de puntos deficiente. En los resultados obtenidos, se observó que, en los frames sin procesar, ciertos puntos clave, como los de las muñecas, presentaban una calidad muy baja o eran indetectables debido a la falta de información o a valores de confianza reducidos. Tras aplicar técnicas de interpolación inspiradas en el modelo G2P-DDM, se logró reconstruir y suavizar las trayectorias de estos puntos, obteniendo coordenadas más precisas y coherentes. Esta mejora permitió una representación más fluida y realista de las poses, especialmente en secuencias donde la calidad original era deficiente. Sin embargo, no se aplicó esta técnica de forma sistemática sobre el dataset principal (RTVE), ya que los puntos faltantes o de baja confianza eran escasos, y el coste computacional añadido no justificaba su uso generalizado.

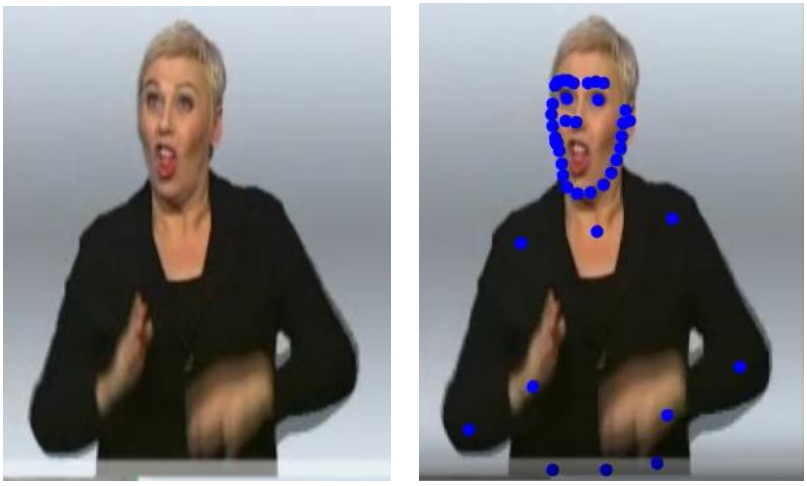

Fig. 10. Ejemplo de extracción de puntos con G2P-DDM.

2.6.4.5 Filtrado defFrames con poca confianza (Pruning)

Explicación corta:

Inspirado en G2P-DDM, se aplicó un método de pruning (poda) para eliminar los frames cuya calidad era baja, debido a puntos clave con poca confianza, limpiando así el conjunto de datos.

Explicación más detallada:

Durante el análisis de G2P-DDM se identificó un enfoque para mejorar la calidad de los datos basado en el filtrado (pruning) de frames. Este método consiste en el cálculo de la confianza promedio de los puntos clave seleccionados (hombros o manos) y eliminar aquellos frames cuyo valor esté por debajo de un umbral predefinido. Aunque esta técnica ayudaba a reducir el ruido en los datos, también producía interrupciones en la secuencia gestual, afectando negativamente a la continuidad del movimiento. Por esta razón, se priorizó el uso de técnicas de interpolación para corregir los frames

problemáticos en lugar de eliminarlos, conservando así la coherencia temporal del gesto y evitando vacíos que pudieran perjudicar el entrenamiento.

Resultados:

La aplicación de pruning permitió filtrar el conjunto de datos eliminando frames con baja confianza en los puntos clave. Sin embargo, este proceso introdujo discontinuidades en los gestos, ya que la supresión de frames intermedios rompía la fluidez natural del movimiento. Como consecuencia, aunque la calidad individual de los frames restantes era mayor, la coherencia temporal de las secuencias se vio comprometida, lo que limitaba la utilidad de esta técnica en contextos de entrenamiento continuo.