Asistente virtual para personas tetrapléjicas controlado por voz y gestos faciales (Primera Fase)

José HERNANDO MosquerA De LA Cruz

GRUPO DE INVESTIGACIÓN KHIMERA

Facultad de Ciencias Básicas e Ingeniería

Fundación Universitaria Católica Lumen Gentium - UNICATÓLICA

Investigación terminada

INTRODUCCIÓN

Desde su creación, los computadores y aplicaciones fueron concebidos para personas sin limitaciones psicofisiológicas y con formas de interacción poco naturales (mouse y teclado), las cuales prevalecen hasta el día de hoy, por esta razón, se hace necesario utilizar mecanismos de interacción natural entre las personas y los computadores mediante el flujo bidireccional de información como el lenguaje hablado y corporal, posibilitando el acceso a la información de forma natural a cualquier persona sana y abriendo un abanico de nuevas oportunidades para personas tetrapléjicas que han sufrido un trauma raquimedular a nivel cervical entre C3 y C7, quedando sin control pleno o parcial de las extremidades superiores e inferiores de su cuerpo, pero con desempeño cognitivo normal (Faroom, et al., 2008).

Con base en lo anterior, este proyecto de investigación desarrolló la primera fase de un asistente virtual para personas tetrapléjicas, el cual interpreta gestos faciales mediante la captura de vídeo y su procesamiento con técnicas de visión artificial. Al mismo tiempo, el asistente virtual identifica comandos de voz, independientemente del locutor y los procesa utilizando técnicas de inteligencia

computacional estableciendo un diálogo multimodal con el usuario, permitiéndole comandar la aplicación WhatsApp Desktop sin tener contacto físico con el mouse y el teclado (Mosquera, et al.,2017).

PLANTEAMIENTO DEL PROBLEMA Y JUSTIFICACIÓN

El desempeño de los equipos de cómputo ha mejorado exponencialmente a la par con los avances en la ingeniería electrónica. Sin embargo, los periféricos de interacción existentes entre las personas y los computadores no han sufrido un avance significativo, pues aún se conservan con ligeras modificaciones el teclado y el ratón como principal medio de interacción. En el caso de las personas tetrapléjicas el computador se convierte en un elemento clave de inclusión, ya que permite la interacción con la información y con otras personas. En estos casos los asistentes virtuales han sido la mejor elección para lograr que los pacientes puedan generar órdenes al computador basados en el reconocimiento de voz (Mertl, et al., 2017; Rani, et al., 2018).

Los asistentes virtuales, basados en voz, orientados a la manipulación del computador por parte de personas con limitaciones motrices, como los pacientes con lesiones medulares a nivel cervical, presentan varios desafíos. El primero de ellos está relacionado con la adaptación de los asistentes virtuales a las nuevas formas de pronunciación de los fonemas por parte de los pacientes con tetraplejía. En la mayoría de los casos los pacientes deben contar con una traqueotomía que ayuda a la respiración cuando el camino habitual presenta alguna complicación, sin embargo, este procedimiento quirúrgico modifica el flujo de aire a través de las cuerdas bucales cambiando las características frecuenciales y de amplitud durante el habla.

Un segundo desafío se encuentra en la calibración o configuración de los asistentes virtuales. La calibración o configuración de los movimientos del cursor deben permitir una personalización rápida y con la menor cantidad de datos posibles; desafortunadamente en el caso de los pacientes tetrapléjicos la manipulación precisa del cursor a partir de la voz, lo cual es un gran reto, debido a la gran cantidad de datos requeridos para generar posiciones efectivas. Una calibración supervisada de los movimientos del cursor requiere muchos datos, por esto es necesario concebir asistentes virtuales que calibren con la menor cantidad de datos posible y en el mejor de los casos de forma no supervisada.

Un último desafío identificado durante la revisión de la literatura está relacionado con la poca retroalimentación que existe entre el paciente y el sistema de inteligencia detrás de los asistentes virtuales. La mayoría de los asistentes virtuales disponibles a nivel comercial no permiten corregir o dar cuenta de falsos positivos o verdaderos negativos. Se hace necesario concebir un asistente que se comunique con el paciente para generar en línea muestras que, al ser utilizadas nuevamente para calibración, permitan mejorar la precisión, por ejemplo, que el asistente pregunte al paciente sobre la veracidad en la decodificación de un comando de voz cuando éste presente una baja probabilidad de certeza (Rani, et al., 2018; Basharirad & Moradhaseli, 2017).

Consideramos que una adecuada fusión de información gestual (proveniente de la cabeza) y de voz de los pacientes permitirá la generación de comandos efectivos en la manipulación de computadores. Con base en lo anterior, en este proyecto se define el problema como: ¿Cuál es la arquitectura de un asistente virtual que permita realizar las tareas básicas de manipulación de una aplicación computacional utilizando gestos corporales y comandos de voz capturados por un sistema cámara – micrófono?

OBJETIVOS

Objetivo general

Desarrollar la primera fase de un asistente virtual que permita a personas tetrapléjicas realizar tareas básicas de manipulación de una aplicación computacional orientada a la navegación en Internet utilizando gestos faciales y comandos de voz mediante el procesamiento de señales de audio y video.

Objetivos específicos

- Determinar las principales características de los asistentes virtuales comandados por señales de audio y vídeo.
- Especificar e implementar un conjunto de técnicas para controlar una aplicación computacional utilizando comandos de voz.
- Especificar e implementar un conjunto de técnicas para controlar una aplicación computacional utilizando gestos faciales.
- Establecer un protocolo de pruebas que permita definir los alcances y limitaciones de las técnicas de interacción vocal y gestual implementadas.

REFERENTE TEÓRICO

A continuación, se presentan los apartados teóricos y conceptuales que enmarcan el desarrollo de este proyecto de investigación:

Interacción humano-máquina

La interacción que se lleva a cabo entre humano-máquina es una doctrina que se enfoca en definir sistemas computacionales que apoyan las tareas diarias del ser humano; el mundo de la tecnología avanza muy rápido, por lo tanto, trae consigo cambios importantes con los cuales se han podido desarrollar varias formas de interacción, tales como interfaces táctiles, hápticas, gestuales y de reconocimiento de voz. En otras palabras, la interacción humano-máquina brinda al usuario una forma de interactuar más natural y novedosa a diferencia del formato clásico como el mouse y teclado (Billard, et al., 2016).

Interacción verbal

Los sistemas de interacción verbal corresponden al proceso de capturar señales de voz que se recogen a través de un micrófono, y convertirlas a texto. Este procesamiento puede ser dependiente o independiente del locutor; el primero hace referencia a un sistema que únicamente reconoce comandos pronunciados por un usuario, específicamente, gracias a un entrenamiento que extrae caracte- rísticas frecuenciales que buscan modelar el tracto vocal del usuario en cuestión. Por otro lado, los sistemas de reconocimiento independiente del locutor funcionan para cualquier usuario, extrayendo características frecuenciales que clasifican la formación de los fonemas pronunciados por cualquier usuario.

Interacción gestual

La IHC gestual es una rama de la IHC que, junto con la lingüística y las ciencias de la computación, tienen como objetivo interpretar gestos humanos a través de algoritmos matemáticos. Los gestos pueden ser cualquier movimiento corporal o estado, pero comúnmente se originan a partir de la cara o la mano. Enfoques actuales en el campo incluyen reconocimiento de la emoción facial y el reconocimiento de gestos de la mano. Muchos enfoques que se han hecho hacen uso de cámaras y algoritmos para interpretar el lenguaje de señas. El reconocimiento de gestos puede ser visto como una manera para que las computadoras empiecen a entender el len-

guaje corporal humano, construyendo así una relación más sólida entre máquinas y seres humanos, dejando atrás sistemas primitivos como las interfaces de usuario de texto o incluso GUIs (interfaces gráficas de usuario), que aún limitan la mayoría de las entradas informáticas al teclado y al ratón. Dado que esta propuesta se enfoca en personas con limitación de miembros superiores, a continuación, se consideran las IHC basadas en gestos faciales.

Acceso a las tics para personas tetrapléjicas

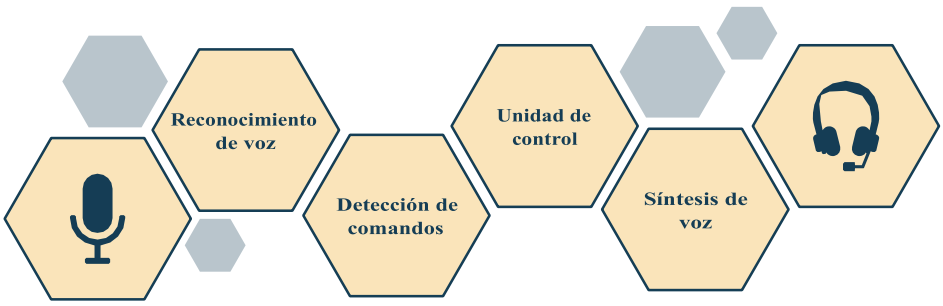
El desarrollo de este proyecto de investigación es fundamental para mejorar la calidad de vida de las personas tetrapléjicas, la cuales han perdido una de las partes más importantes de la vida "la autonomía", la libertad de poder realizar acciones valiéndose por sí mismas y poder acceder a la información del mundo mediante Internet.

METODOLOGÍA

Sistema de interacción verbal

La Figura 1 presenta el diagrama de bloques de la interfaz propuesta. El primer bloque representa el reconocimiento de voz que se encarga de convertir la señal de voz en una cadena de texto. El segundo detecta si la cadena de texto contiene o no los comandos establecidos para el manejo de la aplicación, de lo contrario, la interfaz asimila que es un texto libre que se activa durante un dictado. En el bloque de unida de control se ejecuta el comando de forma automática en el sistema operativo, y finalmente el bloque de síntesis de voz genera una realimentación al usuario que confirma el comando identificado y ejecutado.

Figura 1. *Diagrama de bloques sistema de procesamiento de voz*

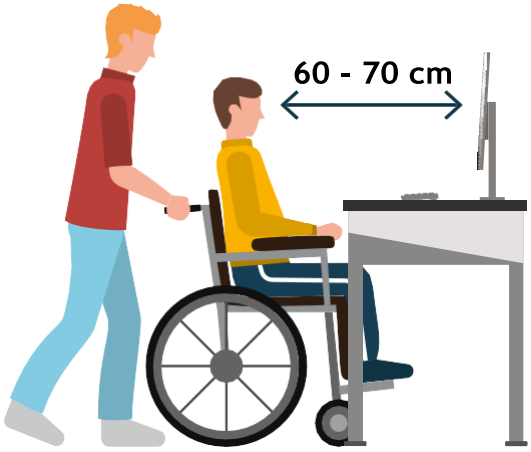


Fuente: elaboración propia.

Sistema de interacción gestual

Para concebir la solución de interfaz gestual primero se definió el software WhatsApp de Escritorio como aplicación objetivo para comandar, esto con base en una encuesta preliminar sobre la preferencia de aplicación de las personas con limitaciones motrices para comunicarse con el mundo exterior. Las condiciones de interacción se establecen en la Figura 2. Para tener una buena composición en la imagen capturada mediante la cámara integrada en un laptop sobre una mesa, se define una distancia promedio entre 60 a 70 cm medidos desde la pantalla del computador, esta debe estar inclinada no más de 20° respecto a la normal de la superficie de apoyo del computador. La primera opción para interactuar con WhatsApp es la de controlar la posición del cursor con ayuda de los movimientos del rostro, dado que estos pueden ser estimados utilizando una técnica de detección rostro; sin embargo, esto puede resultar tedioso para una persona con limitación motriz, ya que debe estar moviendo constantemente la cabeza para seleccionar con el cursor las zonas de interés.

Figura 2. Condiciones de disposición geométrica para interacción con la interfaz gestual



Fuente: ELABORACIÓN PROPIA.

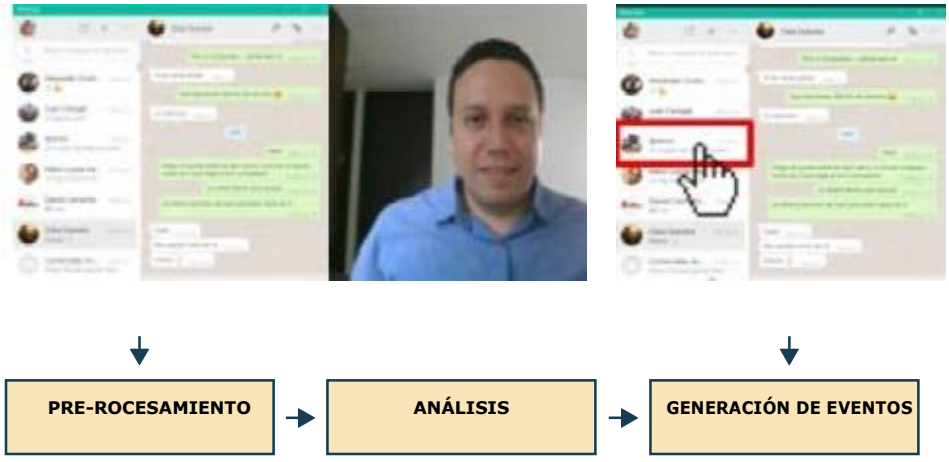
Para hacer más fácil la interacción se pro[illegible]plazamientos específicos del cursor aprovechando que en la ap[illegible]en dos zonas de interacción claramente diferenciadas: z[illegible]ona de contactos, y la zona de chat (o conversación). Por esta razón se propone n[illegible]o solamente procesar la imagen del rostro capturada mediante una webcam sino [illegible]también la captura de pantalla del computador donde la persona está visualizando WhatsApp.

dfpena
2022-05-17 15:43:10
---
alineacion a la izquierda

Adicionalmente se propone como comandos de interacción: i) activar o desactivar la interfaz gestual con la apertura de la boca, ii) ubicar el cursor en la zona de conversación o en la zona de contactos mediante movimientos laterales y sostenidos del rostro, iii) hacer scroll up/down en cualquiera de estas zonas mediante movimientos verticales y sostenidos del rostro. El efecto de hacer scroll up/down en la zona de conversación es la de permitir subir y bajar en la conversación que se ha desarrollado entre dos personas o en un grupo personas. En caso de que el scroll up/down se realice en la zona de contactos permitirá seleccionar un con- tacto a la vez, esto implica emular 'Ctrl+TAB' o 'Ctrl+Shift+TAB' cada que vez que el cursor se desplaza de forma discreta entre los diferentes contactos.

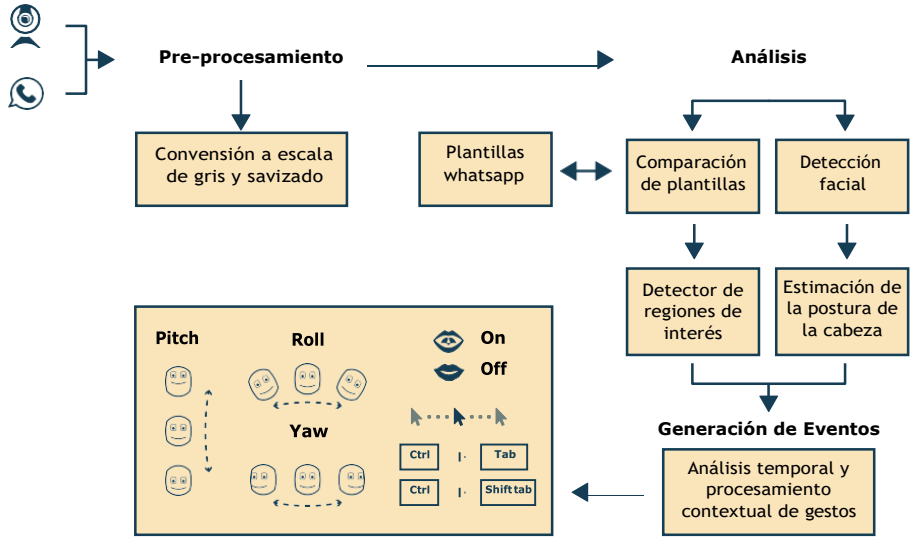
La solución propuesta contempla tres etapas: pre-procesamiento, análisis y generación de eventos, (ver Figura 3). Las técnicas utilizadas en cada una de estas etapas se pueden ver en detalle en la Figura 4.

Figura 3. *Diagrama de bloques principales de la solución propuesta*



Fuente: elaboración propia.

Figura 4. Técnicas implementadas en cada una de las etapas de la solución propuesta



Fuente: ELABORACIÓN PROPIA.

CONCLUSIONES

Se desarrolló una interfaz, la cual permite a personas con limitaciones motrices o tetraplejía comandar la aplicación WhatsApp Desktop mediante comandos de voz y gestos faciales capturados por un conjunto de dispositivos cámara-micrófono, emulando las acciones del mouse-teclado y permitiendo una interacción libre de contacto.

El sistema de reconocimiento de voz permitió realizar las tareas de navegación más comunes de la aplicación de WhatsApp: enviar un mensaje, ir a una conversación en específico o crear una nueva conversación. Se realizaron pruebas con un grupo de personas de 7 personas con diferentes edades y hábitos sobre el uso de internet; en las pruebas se utilizaron 672 comandos de voz y 42 dictados, obteniendo un mayor porcentaje de acierto en los dictados con un 93% y un menor porcentaje en los comandos con un 91.76% de aciertos.

En los comandos de voz se observó que las palabras con estructura fonética similar presentan fallos, pues existía la posibilidad de confundir dos comandos entre sí, un ejemplo es “activar” y “desactivar”, que al ser muy similares si el usuario habla antes de que el micrófono se active puede llegar a confundir los coman-

dos, también se observó cómo los que terminaban con "ar" tuvieron más fallos con un 7.70% de las pruebas realizadas.

En los dictados se observó que no se reconocieron los signos de puntua- ción, pero sí se reconoció el texto pronunciado, a su vez se pudo observar cómo existe una relación directamente proporcional en el reconocimiento de palabras en inglés y la pronunciación del usuario, dando evidencia de poca asistencia a un lenguaje diferente al parametrizado.

El sistema de interacción por comandos de voz fue evaluado por personas que no tenían limitaciones motrices. Con base a su experiencia con el aplicativo se concluyó que la herramienta puede mejorar la calidad de vida de personas con limitaciones motrices, pues le permite controlar la aplicación de mensajería sin utilizar el teclado o muse, adicionalmente tener más privacidad al momento e interactuar con otras personas sin asistencia de terceros.

Se desarrolló una interfaz humano-computador basada en gestos faciales y detección de zonas de interés en una aplicación orientada al internet para personas con limitaciones motrices de miembros superiores. La interfaz permite, mediante cinco comandos, seleccionar contactos, navegar en las conversaciones de la aplicación WhatsApp de escritor, así como y activar/desactivar la interfaz utilizando únicamente gestos del rostro.

Se determinaron las principales características de una interfaz huma- no-computador para personas con limitaciones motrices de miembros superiores. La interfaz cuenta con principios ergonómicos para facilitar y optimizar su utilización. También posee una combinación de tecnología, conocimiento y recursos que produjeron resultados deseados en equipos de cómputo con precios inferiores a los 1.5 millones pesos colombianos.

Se implementó una técnica de visión por computador para la identifica- ción de gestos faciales y la detección de zonas de interés en una aplicación de escri- torio orientada a internet. Se logró la detección automática a partir de una captura de pantalla de la aplicación WhatsApp, las zonas de contacto y conversaciones mediante una combinación de heurísticas y correspondencia por plantilla de imá- genes. La detección de gestos se logró utilizando técnicas robustas de detección de rostros y estimación de pose mediante modelos flexibles.

Se construyó una interfaz software de generación de comandos para una aplicación de escritorio a partir de gestos y zonas de interés detectadas. Esta interfaz es capaz de correr en fondo y no interfiere visualmente con la aplicación WhatsApp. La velocidad de ejecución de la misma es de 1 Hz, y en un equipo de media gama ocupa el 35 % de CPU y 1.5 GB RAM.

Se llevaron a cabo pruebas de desempeño computacional, robustez y de usabilidad. Primero, las pruebas de desempeño permitieron identificar que en un equipo de media gama es posible trabajar la interfaz gestual y que se pueden tener más aplicaciones ejecutándose simultáneamente sin que esto reduzca la velocidad de ejecución de esta. Segundo, las pruebas de robustez evidenciaron que la interfaz gestual puede trabajar en variadas condiciones de iluminación y con algunos casos de oclusión, tanto a nivel de detección de gestos faciales como detección de zonas de interés. Por último, las pruebas de usabilidad permitieron entender que los usuarios reconocen la funcionalidad de la interfaz implementada, así como su potencial para personas con limitación motriz, sin embargo, sugieren que debe ampliarse el número de eventos y de capacidades de la interfaz para ser completamente útil, además de sugerir una mejora en los tiempos de ejecución para que el manejo sea más fluido.

En general, los autores de este trabajo consideran que los hallazgos encontrados en la ejecución de este proyecto son claves para su aplicación en personas con limitaciones motrices de miembro superior y que deben integrarse a la parte de reconocimiento de voz (el cual es otro proyecto que se está realizando en la Fundación Universitaria Lumen Gentium) para complementar su espectro de aplicación.

Finalmente, con el fin de ampliar el alcance obtenido y de igual forma superar las limitaciones de la solución propuesta en este trabajo de investigación se propone a futuro i) incluir más eventos a partir de la identificación de otros gestos del rostro, por ejemplo, los que involucran los ojos ii) Mejorar el tiempo de ejecución de la interfaz gestual mediante la re-implementación de algunas técnicas de análisis de imágenes para que escalen mejor las capacidades de la CPU e iii) integrar reconocimiento por voz para permitir el ingreso de texto utilizando técnicas de procesamiento del lenguaje natural.

REFERENCIAS

Basharirad, B. and Moradhaseli, M. (2017). *Speech EMOTION RECOGNITION Methods: A LITERATURE. Review.* 2nd Int. Conf. Appl. Sci. Technol. doi: 10.1063/1.5005438.

Billard, A. G., Calinon, S. and Dillmann, R. (2016). Learning from Humans. Springer Handb. Robot, p.p. 1995-2014.

Estes, S., Helleberg, J., Long, K., Pollack, M. and Quezada, M. (2008). *GUIDELINES for speech INTERACTIONS BETWEEN pilot & COGNITIVE ASSISTANT.* ICNS 2018 - Integr. Commun. Navig. Surveill. Conf., pp. 1–23. Doi: 10.1109/ICNSURV.2018.8384965.

Faroom, S., Nauman, M. Yousaf, S. and Umer. S. (2008). *LITERATURE Review ON Home AUTOMATION System.* Int. Conf. Comput. Math. Eng. Technoldoi: 10.17148/IJARCCE.2017.63173.

Khan, N. H., Arovi, A. H., Mahmud, H. Hasan, K. and Rubaiyeat, H. A. (2015). *Speech BASED text CORRECTION tool for the VISUALLY IMPAIRED.* pp. 150–155.

Lugano, G. (2017). *VIRTUAL ASSISTANTS AND SELF-DRIVING CARS.* Proc. 15th Int. Conf. ITS Telecommun. ITST, pp. 1–5, doi: 10.1109/ITST.2017.7972192.

Mertl, J., Žá-ková, E. and -epová, B. (2017). Quality of life of patients after total laryngectomy: the struggle against stigmatization and social exclusion using speech synthesis. *DISABILITY AND REHABILITATION: Assistive TECHNOLOGY, 13*(4), 1-11.

Mosquera, J.H., Loaiza, H., Nope, S. E. and Restrepo, A. D. (2017). Identifying facial gestures to emulate a mouse: navigation application on Facebook. *IEEE LATIN AMERICA TRANSACTIONS, 15*, pp. 121–128.

Rani, P. J., Bakthakumar, J., Kumaar, B. P., Kumaar, U. P. and Kumar, S. (2018). *Voice CONTROLLED home AUTOMATION system USING NATURAL LANGUAGE PROCESSING (NLP) AND INTERNET of THINGS (IoT).* ICONSTEM 2017 - Proc. 3rd IEEE Int. Conf. Sci. Technol. Eng. Manag., vol. 2018-Janua, pp. 368–373, doi: 10.1109/ICONSTEM.2017.8261311.

Vasco, N; Pinto, F.; Furtado, P. and Silva, J. (2014). *IoT ARCHITECTURE PROPOSAL for DISABLED people.* Int. Conf. Wirel. Mob. Comput. Netw. Commun. pp. 152–158. Doi: 10.1109/WiMOB.2014.6962164.

COMENTARIOS DEL EVALUADOR

Evaluador: Steve Rodríguez Guerrero

Quería conocer el contexto de la presentación porque como el formato de la Unicatólica es muy sencillo, entonces deja por fuera muchas cosas que en la presentación sí se vieron, pero en el formato, muchas veces no quedaron contempladas. Entonces, mis observaciones van al documento, que fue el que, en primera instancia, tuve con relación al ejercicio. Quiero hacer una observación y es que, si bien en el planteamiento del problema haces una descripción de que están definiendo una arquitectura para poder llevar a cabo un asistente, una arquitectura para un asistente que permita hacer reconocimiento de voz y gestos para apoyar el tema de estas personas con tetraplejia, no hay una articulación entre el tratamiento de problema que habla sobre la arquitectura y el objetivo general, que ya se refiere es a otra cosa, o sea no tiene una referencia, no hay una articulación clara entre la pregunta problema y el objetivo general. Entonces yo esperaba que el objetivo general, pues si la pregunta problema me estaba diciendo cuál es la arquitectura que voy a definir para apoyar el tema del reconocimiento, entonces el objetivo podría ir orientado más a una arquitectura o identificar la arquitectura.

Exacto, entonces se separan porque finalmente en el objetivo dices: esta es una primera fase efectivamente; pero entonces, cambia la pregunta. La pregunta problema en la descripción del problema sí habla de que es una arquitectura general, pero la pregunta problema de esa primera fase es totalmente diferente.

Evaluador: Es como si la pregunta problema, esa primera fase es otra diferente a la arquitectura, porque entiendo yo que la arquitectura ya es el macro.

Evaluador: Hay otra observación, en el objetivo general, es que das una orientación a que la aplicación va a ser orientada en Internet y lo que pasa es que, finalmente, cuando termino viendo el desarrollo metodológico, terminamos cayendo en un chat de WhatsApp, entonces ¿la orientación a internet dónde me queda?

Evaluador: Pues digamos que la metodología, cuando haces una descripción metodológica hablas muy general, porque se llega a explicar acerca de los resultados obtenidos y no se habla mucho. Lo que siempre hemos dicho en la parte de inteligencia artificial, que siempre nos han criticado, es que no se puede decir que eso es

una caja negra, o sea, que lo que está allá es una caja negra y no podemos hacer una referencia acerca de cómo se hizo el análisis, cuáles fueron las características que se obtuvieron, porque finalmente eso es lo que le da un peso a la investigación, o sea, qué tipo de análisis realmente hiciste, qué características obtuviste en la voz, en qué frecuencia. Al final lo que obtuve fue el resultado diciéndome: "el noventa por ciento fue fiable en palabras de este sentido, no se correspondió, hubo dificultades", esos resultados están bien, pero en la metodología de pronto había que dar una razón más clara acerca de si en la etapa de análisis se hizo, qué tipo de análisis se hizo...¿variado?

Evaluador: Bueno, aquí hay otra, no tan técnica, sino más orientada a la parte de apoyo. Realmente con cuántos pacientes, en este momento se han hecho pruebas con siete personas.

Evaluador: Basado en la experiencia, hay una cosa que siempre nos sucede y es que la herramienta que estamos desarrollando puede que sea muy avanzada, muy bien elaborada, pero, cuando vamos al campo con el paciente, la herramienta termina siendo rechazada porque la curva de aprendizaje es muy alta.

Evaluador: Finalmente, aquí hay una cosa que dijiste en la presentación, pero la quiero aclarar. Dijiste: "Queremos trabajar con software libre", pero finalmente dentro de las herramientas me dijiste que estabas trabajando con Phyton, entonces ahí no sé si es que ya lo hicieron con un software que no era libre y ahorita lo quieren pasar a un software libre o siempre han estado trabajando con software libre.

En el documento están llegando a un análisis de tipo matriz de confusión, porque están identificando dónde hay la certeza, dónde están las no certezas del dispositivo, pero, entonces, para que esa matriz de confusión sea lo suficientemente enriquecedora y válida, tendrán que determinar una población considerable para poder decir si realmente el sistema lo podemos llegar a generalizar. Entonces lo único que, en el documento, que le agregaría en la parte teórica es un poco acerca de: ¿Qué porcentaje de población con este tipo de situación de la salud encontramos en nuestra población?

Evaluador: La muestra que va a participar en el proyecto, que sea significativa para definir que eso que estamos desarrollando, finalmente sí es significativo y puede llegarse a generalizar.