Análisis de *transformers* para el reconocimiento de voz y evaluación de su rendimiento en sistemas de recursos limitados

Cristina Bolaños Peño[1], Jesús Fernández-Bermejo Ruiz[1], Henry Llumiguano Solano[1], Javier Dorado Chaparro[1], Félix Jesús Villanueva Molina[1] y Juan Carlos López López[1]

Resumen— **El uso de *transformers* para la consecución de tareas relacionadas con el procesamiento del lenguaje natural gana popularidad a cada día que pasa. Este hecho también se extiende al campo del procesamiento de la voz. No sólo compiten con las redes neuronales convolucionales tradicionales en rendimiento, si no que, de la mano de librerías sostenidas por la comunidad de la IA, son fáciles de usar directamente en aplicaciones finales. Sin embargo, estos modelos requieren muchos recursos a medida que aumentan sus capacidades y parámetros manejados, normalmente traducidos en mayor precisión. Este trabajo estudia diferentes alternativas para el reconocimiento de voz que empleen *transformers* con enfoque en su rendimiento y precisión en sistemas de recursos limitados, pudiendo así determinar la opción más adecuada para su aplicación en tiempo real en el contexto de la implementación de un agente conversacional en un dispositivo empotrado.**

Palabras clave— **reconocimiento de voz, evaluación, agente conversacional, sistema empotrado**

I. Introducción

La Inteligencia Artificial (IA) se encuentra profundamente integrada en la vida cotidiana de la sociedad actual, casi de forma natural. Este hecho se refleja en los crecientes esfuerzos de la Unión Europea (UE) por regular su uso [1]. Las polifacéticas aplicaciones de la IA abarcan una amplia gama de ámbitos, como la generación de contenidos («IAs generativas»), la detección de objetos a través de imágenes y el reconocimiento de la voz, entre muchos otros. El ritmo de crecimiento de la tecnología y todo aquello relacionado con la IA se erige como ejemplo por excelencia de una incesante necesidad de innovación en este campo.

Dentro del extenso panorama de técnicas y algoritmos de IA y *deep learning* (aprendizaje profundo), el uso de *transformers*, un tipo de arquitectura de red neuronal, en tareas relacionadas con el procesamiento del lenguaje natural (NPL, *Natural Language Processing*) y el procesamiento de la voz ha cobrado popularidad en los últimos años [2]. A diferencia de las redes neuronales convolucionales (CNNs, *Convolutional Neural Network*) tradicionales, los *transformers* poseen una arquitectura única que facilita el modelado de dependencias de largo alcance dentro de las secuencias de entrada, una capacidad crítica para manejar datos secuenciales como las ondas sonoras. Esta innovación arquitectónica ha catapultado a los *transformers* al primer plano, permitiéndoles rivalizar e incluso superar el rendimiento de las CNNs, al menos en este tipo de aplicaciones. Dadas las características de los *transformers* y su auge, estos mejoran su rendimiento en tareas como el reconocimiento automático de voz (ASR, *Automatic Speech Recognition*) y la traducción del habla (SP, *Speech Translation*).

La proliferación de *transformers* en tareas de reconocimiento de voz ha allanado el camino para el desarrollo de agentes conversacionales y asistentes virtuales avanzados, permitiendo que cualquier herramienta *software* adquiera una interfaz más natural e intuitiva con el usuario final. Estos agentes potenciados por IA aprovechan las capacidades de los *transformers* para comprender y responder a consultas, ejecutar órdenes, e incluso entablar un diálogo significativo con los usuarios. No más lejos de la realidad, aplicaciones de este calibre ya están disponibles hoy en día en forma de asistentes virtuales integrados en *smartphones* [3]. La integración del reconocimiento de voz ha revolucionado la forma en la que se interactúa con la tecnología y accedemos a la información disponible.

A pesar de los notables avances logrados en el campo del reconocimiento de voz de la mano de los *transformers*, aún persisten retos, sobre todo en el contexto de los sistemas de recursos limitados. Lamentablemente, a medida que aumentan las capacidades y precisión de estas redes neuronales también aumentan la cantidad de recursos requeridos para su uso. Los requisitos computacionales y de memoria que exigen los *transformers* pueden resultar extenuantes para dispositivos empotrados, lo cual hace necesario explorar nuevos enfoques y optimizadores para que los sistemas de reconocimiento de voz sean más eficientes y escalables en este tipo de entornos.

En los últimos años se han explorado diversas estrategias para mejorar el rendimiento y la eficiencia de las redes neuronales, incluyendo técnicas como la compresión de modelos, cuantización y aceleradores *hardware* [4]. Al optimizar el despliegue de los *transformers*, se pretende lograr un equilibrio entre precisión, rendimiento y utilización de recursos que permita el reconocimiento de voz en tiempo real en una amplia gama de dispositivos.

La integración del reconocimiento de voz en aplicaciones e industrias encierra un inmenso potencial para impulsar la innovación y mejorar la experiencia del usuario, creando así soluciones más intuitivas,

[1]Dpto. de Sistemas y Tecnologías de la Información, Universidad de Castilla-La Mancha, e-mail: Cristina.Bolanos@uclm.es

personalizadas y accesibles. Sin embargo, no hay que desestimar el coste final del sistema, por lo que los dispositivos de recursos limitados cobran gran importancia en relación al despliegue de la solución propuesta.

Este trabajo estudia diferentes alternativas para el reconocimiento de voz que empleen *transformers* con enfoque en su rendimiento y precisión en sistemas de recursos limitados, pudiendo así determinar la opción más adecuada para su aplicación en tiempo real en el contexto de la implementación de un agente conversacional en un dispositivo empotrado.

II. Trabajos relacionados

Dentro de la amplia variedad de alternativas para el reconocimiento de voz disponibles en la literatura se destacan las siguientes: Wav2Vec2, Conformer, WavLM y Whisper.

Wav2Vec2 [5] parte de audio en bruto y lo codifica para luego, y tomando como referencia el modelo de lenguaje BERT [6], enmascara las representaciones latentes de la voz y alimenta la red de *transformer*. Después, el modelo es entrenado empleando codificación comparativa predictiva (CPC, *Contrastive Predictive Coding*). Esta estrategia ha permitido a Wav2Vec2 posicionarse como uno de los modelos más novedosos y potentes. Más adelante se propone el modelo XLS-R [7], el cual está basado en Wav2Vec2 pero cuyo objetivo es la representación *cross-lingual* o «entre lenguas». La Figura 1 muestra la estructura de Wav2Vec2.

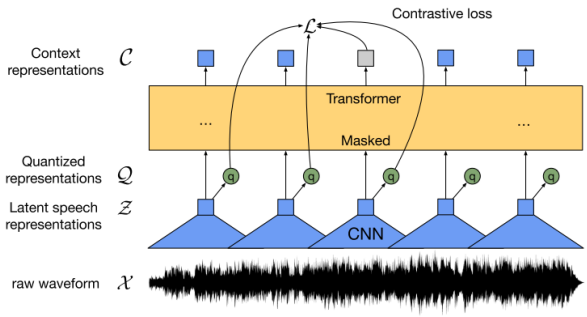

Fig. 1: Estructura de Wav2Vec2 [5].

Conformer [8] emplea una arquitectura que combina CNNs y *transformers*, permitiendo poder capturar información tanto del contexto local como del global. Más adelante se combinó el pre-entrenamiento de Wav2Vec2 con esta arquitectura [9] para obtener buenos resultados, frente a la literatura del momento, en el conjunto de datos LibriSpeech [10]. La Figura 2 muestra la estructura de Conformer.

WavLM [11], basada en la arquitectura de HuBERT [12], se enfoca en la preservación de la identidad del hablante, además de la obtención del contenido textual del audio. WavLM está pre-entrenado en 94k horas de audio de 10 idiomas diferentes y ofrece buenos resultados, frente a la literatura del momento, para las tareas de ASR y la identificación del hablante. La Figura 3 muestra la estructura de WavLM.

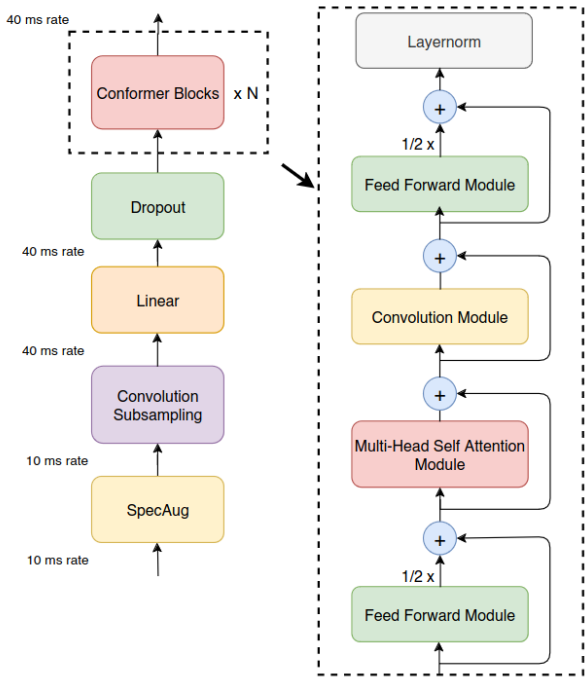

Fig. 2: Estructura de Conformer [8].

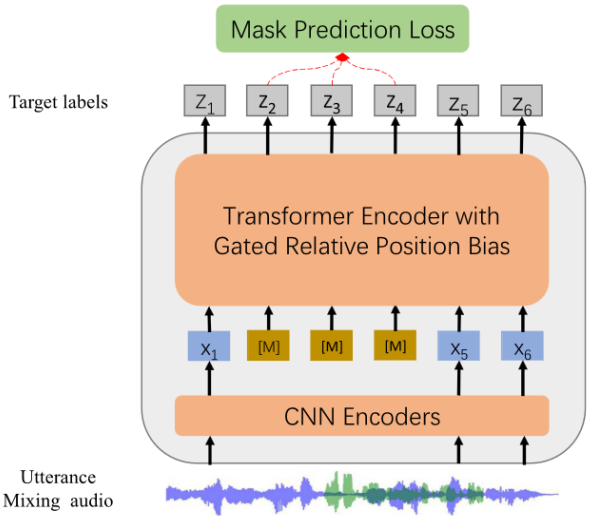

Fig. 3: Estructura de WavLM [11].

Whisper [13] está diseñado para ofrecer grandes resultados en entornos con ruido o con audio de baja calidad, además de ofrecer solución a diferentes tareas relacionadas con el reconocimiento de voz, como el reconocimiento e indentificación de diferentes lenguas y la traducción entre ellas y el inglés. Con un pre-procesamiento de los datos clasificado como «minimalista», Whisper no necesita un paso inverso a la salida del *transformer*, por lo que simplifica el procesamiento del modelo. La última versión de esta alternativa proporciona únicamente un 5 % de error por palabra (WER, *Word Error Rate*) en español. La Figura 4 muestra la estructura de Whisper.

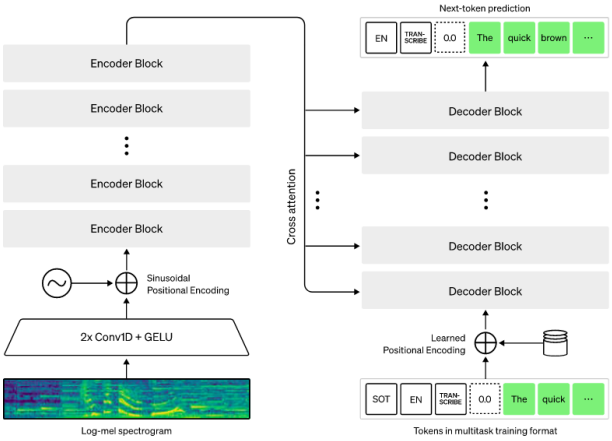

Fig. 4: Estructura de Whisper [14].

III. Métodos

A. Modelos

Este trabajo pretende realizar una comparación exhaustiva de varias alternativas encontradas en la literatura existente que hayan demostrado resultados prometedores en tareas de ASR y ST. La prioridad será evaluar su idoneidad para el despliegue en tiempo real de un agente conversacional hispanohablante dentro de un sistema de recursos limitados.

Es necesario el acceso a versiones ya entrenadas de redes neuronales para poder realizar su procesamiento con datos. La obtención de estos modelos preentrenados se realizará por medio de:

- HuggingFace [15]: Empleando su librería «transformers», entre otras, HuggingFace permite un cómodo acceso a múltiples *transformers* preentrenados de última generación mantenidos por la comunidad del propio portal.
- OpenAI [16]: Empleando su librería «openai-whisper», OpenAI permite el libre acceso a la última versión de su modelo Whisper.

Seleccionando de entre las alternativas expuestas en la Sección II aquellas disponibles mediante las librerías mencionadas y preparadas para el procesamiento del lenguaje castellano, se proponen los siguientes modelos a comparar en este trabajo:

- Wav2Vec2: Este modelo presenta dos variantes:
  - El modelo original de tamaño *base* entrenado en un subconjunto de datos de VoxPopuli [17] en español[1].
  - El modelo en su versión XLS-R (*cross-lingual*) de tamaño *large* entrenado en un subconjunto de datos de Common Voice 6.1 [18] en español[2].
- Conformer: Este modelo presenta una variante de tamaño *large* entrenado en subconjuntos de datos de Common Voice 7.0 y Multilingual LibriSpeech [19] en español[3].
- Whisper: Este modelo presenta cuatro variantes: *tiny*, *small*, *base* y *large*. Todas las variantes de este modelo fueron entrenadas empleando diversos conjuntos de datos[4].

B. Conjunto de datos

Al evaluar un modelo en situaciones reales y cercanas al caso de uso objetivo, lejos de los entornos experimentales controlados o *benchmarks* académicos, se analiza la robustez y adaptabilidad del mismo. Los conjuntos de datos «in the wild» o «no controlados» presentan variabilidad, ruido y escenarios inesperados que pueden influir en le rendimiento global de un algoritmo de IA.

Con el fin de evaluar exhaustivamente la precisión y el rendimiento de los modelos propuestos en este tipo de datos, se ha elaborado un conjunto de datos compuesto por 65 archivos de audio. Estos archivos contienen frases de tamaño variable en español, seleccionadas para simular órdenes que el agente conversacional que se desea desarrollar puede encontrar en escenarios reales. La duración máxima de dichos archivos de audio son 5 segundos, siendo 3.45 segundos la duración media de todos los archivos propuestos. Ejemplos de estas frases son «Quiero hablar con José» o «¿Podrías llamar a Antonio?».

Dichos archivos de audio provienen de las voces de 7 sujetos distintos. La preparación del entorno de grabación de los archivos de audio comparten las características en las que se espera encontrar el agente conversacional final. Factores como el tipo de micrófono y la distancia entre el sujeto y el dispositivo de grabación se controlaron meticulosamente para la simulación de condiciones realistas.

El dispositivo de recursos limitados empleado en esta evaluación se trata de un NiPoGi AK2 Pro. Este económico dispositivo cuenta con una serie de especificaciones, entre las que se incluyen: 12 GB de RAM, una CPU Intel Pentium J4205 y gráficos Intel HD Graphics 505. Con un precio de venta al público de 199,99 euros[5], este dispositivo ofrece una solución asequible y se espera que el agente conversacional pueda operar en tiempo real en este dispositivo o alguno de características similares.

IV. Resultados

Siguiendo los métodos descritos en la Sección III, los modelos propuestos se emplearán para procesar los archivos de audio, obteniendo las transcripciones de cada uno de ellos. Por un lado, se evaluará la calidad de dichas transcripciones en términos de precisión, por otro, se analizará el tiempo requerido para procesar los archivos dentro de un dispositivo de recursos limitados. Finalmente, se seleccionará la mejor alternativa de entre las propuestas para el caso de uso que abarca el presente trabajo.

[1] facebook/wav2vec2-base-10k-voxpopuli-ft-es
[2] facebook/wav2vec2-large-xlsr-53-spanish
[3] nvidia/stt_es_conformer_ctc_large

[4] `https://github.com/openai/whisper/tree/main/data`
[5] Precio orientativo. Disponibilidad desconocida. Consultado el 2024-03-10.

A. Comparación de la precisión

Las transcripciones de los archivos de audio obtenidas de los modelos propuestos se han evaluado tomando en consideración las siguientes métricas [20]:

- WER (*Word Error Rate*): Porcentaje de error por palabra. Es la métrica más comúnmente utilizada. Siendo A, S, I y E los aciertos, sustituciones, inserciones y eliminaciones, respectivamente, y N_1 y N_2 el número de elementos de las secuencias de entrada y salida del modelo, respectivamente, este ratio es calculado de la forma:
$$WER = \frac{S + I + E}{N_1}. \quad (1)$$
- CER (*Character Error Rate*): Porcentaje de error por caracter. Comparte fórmula con WER, pero los elementos de las secuencias de entrada y salida no son palabras, si no los caracteres que las componen.
- WIP (*Word Information Preserved*): Porcentaje de información preservada por palabra. Siendo A, S, I y E los aciertos, sustituciones, inserciones y eliminaciones, respectivamente, y N_1 y N_2 el número de elementos de las secuencias de entrada y salida del modelo, respectivamente, este ratio es calculado de la forma:
$$WIP = \frac{A}{N_1}\frac{A}{N_2}. \quad (2)$$

La Figura 5 muestra los resultados de la evaluación de los modelos propuestos frente a la métrica WER. Ordenados de menor a mayor error obtenido: Whisper *large* (11 %), Conformer (17.7 %), Whisper *small* (24.63 %), Wav2Vec2 XLS-R *large* (25 %), Whisper *base* (43.95 %), Wav2Vec2 *base* (45.96 %) y Whisper *tiny* (60 %).

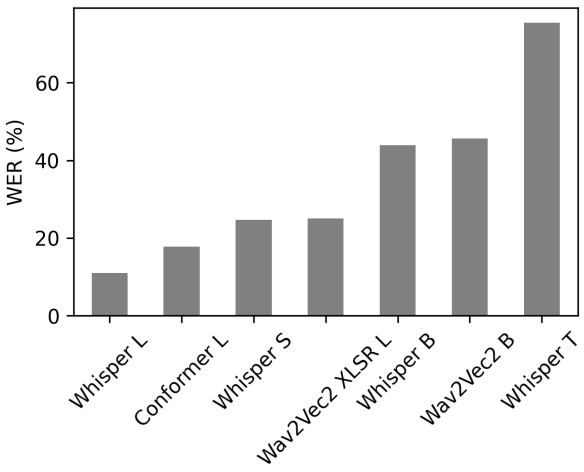

Fig. 5: Ratio de error por palabra (WER).

En la Figura 6 pueden apreciarse los resultados de la evaluación frente a la métrica CER. Ordenados de menor a mayor error obtenido: Whisper *large* (5.7 %), Conformer (6.3 %), Wav2Vec2 XLS-R *large* (7.2 %), Whisper *small* (12.38 %), Wav2Vec2 *base* (12.41 %), Whisper *base* (22.54 %) y Whisper *tiny* (29.9 %).

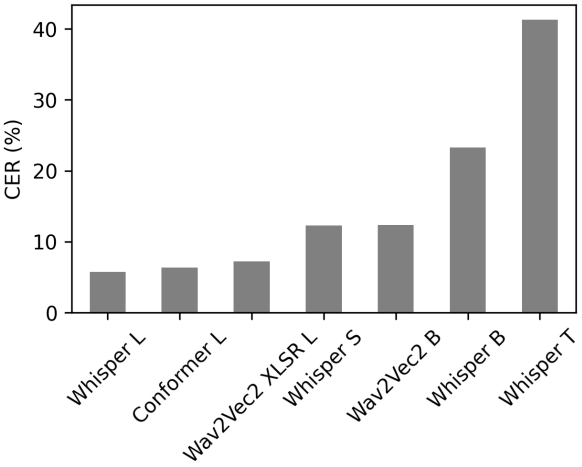

Fig. 6: Ratio de error por caracter (CER).

En la Figura 7 se ven reflejados los resultados de la evaluación frente a la métrica WIP. Para esta métrica cuanto mayor es el resultado, mejor se le considera, al contrario que las anteriores métricas. Ordenados de mayor a menor información preservada: Whisper *large* (84.94 %), Conformer (75 %), Whisper *small* (68.5 %), Wav2Vec2 XLS-R *large* (66.46 %), Whisper *base* (48.33 %), Wav2Vec2 *base* (43.6 %) y Whisper *tiny* (34.4 %).

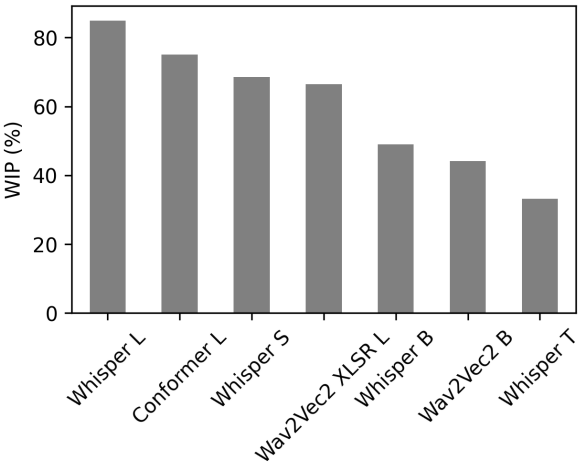

Fig. 7: Ratio de información preservada por palabra (WIP).

Parece existir una correlación entre el tamaño del modelo (*large*, *small*, *base* y *tiny*) y la precisión obtenida, ya que los modelos grandes suelen ofrecer mejores resultados. A mayor número de parámetros manejados por la red neuronal, mayor es la precisión que ofrece.

En vista de los resultados, el modelo Whisper, en todas sus variantes de forma general, obtiene peores resultados al ser evaluado frente a la métrica CER. CER provee de una vista más detallada de la transcripción, al evaluar caracter a caracter.

Whisper *large* destaca como la mejor alternativa para el reconocimiento de voz en relación a su precisión evaluada para los archivos de audio propuestos.

B. Comparación del rendimiento

Al procesar cada archivo de audio, los segundos empleados en esta tarea fueron anotados. La Figura 8 muestra el tiempo medio que emplean cada uno de los modelos propuestos en procesar un archivo de audio. Ordenados de menor a mayor tiempo de inferencia: Conformer (2.86s), Wav2Vec2 *base* (3.83s), Wav2Vec2 XLS-R *large* (9.42s), Whisper *tiny* (14s), Whisper *base* (26.63s) y Whisper *small* (58s).

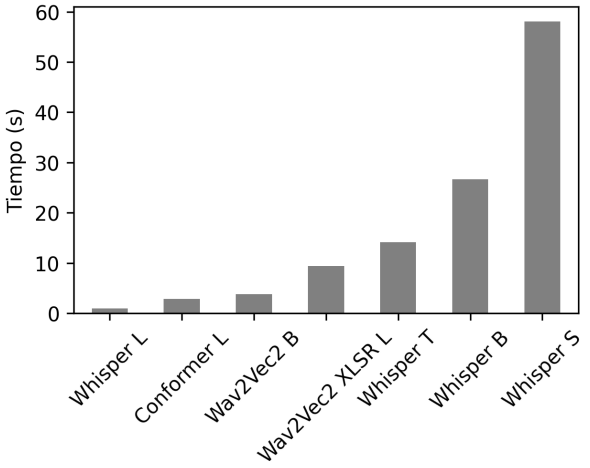

Fig. 8: Tiempo de inferencia medio para archivos de audio de entre 3 y 5 segundos de duración.

El modelo Conformer obtiene la primera posición con **2.86s** de media por archivo de audio, seguido por Wav2Vec2 *base* y Wav2Vec2 XLS-R *large* con 3.83s y 9.42s, respectivamente. Se observa una correlación entre el tamaño del modelo y la velocidad de procesamiento, siendo más rápidos los modelos más pequeños.

De forma general, el modelo Whisper se ve gravemente penalizado por el tiempo de inferencia frente a otros modelos. Whisper *large* desaparece del gráfico debido a la imposibilidad de ejecutarlo por la insuficiencia de memoria RAM disponible en el dispositivo utilizado.

Conformer destaca como la mejor alternativa para el reconocimiento de voz en relación a su rendimiento evaluado para los archivos de audio propuestos.

C. Selección

Una vez obtenidos los resultados de la evaluación, los modelos han de ser clasificados según la idoneidad para el caso de uso propuesto. Dichos modelos ya han sido clasificados para cada métrica detallada en las Secciones IV-A y IV-B.

Si el objetivo de un asistente virtual fuera desarrollarse única y exclusivamente para recibir órdenes o comandos por parte del usuario, la precisión del modelo utilizado en cuestión debe ser máxima, independientemente del rendimiento temporal. Sin embargo, en el caso del desarrollo de un agente conversacional más equilibrado, debe emplearse una forma de valoración que incluya tanto la precisión como el rendimiento.

El objetivo de este trabajo es encontrar el mejor modelo, no el más preciso. La puntuación final de un modelo consistirá en la suma de las posiciones que haya ostentado en las clasificaciones tanto de precisión como de rendimiento. Siendo P una puntuación, la puntuación de un modelo es calculada de la siguiente forma:

$$P_{total} = P_{wer} + P_{cer} + P_{wip} + P_{tiempo} \qquad (3)$$

El modelo que **menor** puntuación obtenga se considerará el mejor de entre los propuestos. En la Tabla I se muestran las puntuaciones finales para cada modelo.

Tabla I: Puntuación de los modelos propuestos, ordenados de menor a mayor. Los modelos corresponden a: Conformer (1), Wav2Vec2 XLS-R *large* (2), Whisper *small* (3), Wav2Vec2 *base* (4), Whisper *base* (5), Whisper *tiny* (6) y Whisper *large* (7).

	Posiciones				
Modelo	**WER**	**CER**	**WIP**	**Tiempo**	**Puntuación**
1	2	2	2	1	7
2	4	3	4	3	14
3	3	4	3	6	16
4	6	5	6	2	19
5	5	6	5	5	21
6	7	7	7	4	25
7	1	1	1	∞	∞

Según el mecanismo de puntuación propuesto, Conformer se distingue como el mejor modelo para el reconocimiento de voz en el caso de uso propuesto en el presente trabajo.

V. Conclusiones

Este trabajo expone varios modelos para el reconocimiento de voz para su aplicación en tiempo real en el contexto de la implementación de un agente conversacional en un dispositivo empotrado, destacando de entre ellos Conformer como la solución más óptima. Sin embargo, la continua evolución de este campo del conocimiento sugiere que es inevitable que se produzcan nuevos avances.

El español, como cualquier otro idioma, posee un rico repertorio de acentos y dialectos. Si bien los modelos existentes funcionan bien con acentos estándar, suelen tener dificultades al enfrentarse a variaciones no estándar y/o regionales. Afrontar este reto promete reforzar la inclusión y accesibilidad, garantizando que los sistemas de reconocimiento de voz se adapten eficazmente a los diversos grupos poblacionales de usuarios.

El ajuste individualizado a cada usuario representa una vía prometedora para mejorar la precisión y personalización de un sistema de reconocimiento de voz. Al permitir a los usuarios entrenar los modelos con sus propios patrones de habla y características de voz, estos sistemas pueden adaptarse mejor a las voces individuales, lo que se traduce en una mayor precisión y satisfacción del usuario. Este enfoque puede implicar el aprovechamiento de técnicas como el aprendizaje por transferencia y el *fine-tunning* para adaptar los sistemas a los rasgos únicos y específicos de cada usuario.

Además, en un diálogo se presentan formas de comunicación verbales y no verbales. La integración de información visual y contextual ofrece un enfoque prometedor para mejorar la comprensión y las capacidades de los sistemas de reconocimiento de voz. Estos modelos multimodales tienen el potencial de mejorar la precisión, esclarecer el significado del discurso y comprender la intención del usuario, facilitando así interacciones aún más naturales e intuitivas.

En esencia, la trayectoria futura del reconocimiento del habla pasa por aumentar su adaptabilidad, precisión y robustez para ajustarse mejor a las necesidades del usuario. Al abordar retos como la variación del acento, la individualización y el rendimiento en tiempo real, el potencial transformador de esta tecnología puede allanar el camino para su adopción generalizada en diversos sectores y aplicaciones.

Agradecimientos

El presente trabajo ha sido financiado por:

- Proyecto MIRATAR (TED2021-132149B-C42), financiado por el Ministerio de Ciencia e Innovación de España y por la Unión Europea con los fondos NextGeneration.
- Proyecto TALENT-BELIEF (PID2020-116417RB-C44), financiado por MCIN/ AEI /10.13039/501100011033.
- 2023-UNIVERS-11977-FPU08 Formación de Profesorado Universitario (FPU) financiado por el Programa Fondo Social Europeo Plus (FSE+).

Referencias

[1] Council of the EU, "Artificial intelligence act: Council and parliament strike a deal on the first rules for ai in the world," `https://europa.eu/!hjMd9Q`, 2024, Accedido en 2024-03-10.

[2] S. Latif, A. Zaidi, H. Cuayáhuitl, Fahad Shamshad, Moazzam Shoukat, and Junaid Qadir, "Transformers in speech processing: A survey," *arXiv.org*, 2023.

[3] Abraham Andreu, "Historia de los asistentes virtuales como alexa, siri y google, y el principio del fin ante el auge de la ia tipo chatgpt," `https://computerhoy.com/android/historia-siri-alexa-otros-asistentes-futuro-chatgpt-1237612`, 2023, Accedido en 2024-03-10.

[4] Lei Deng, Guoqi Li, Song Han, Luping Shi, and Yuan Xie, "Model compression and hardware acceleration for neural networks: A comprehensive survey," *Proceedings of the IEEE*, vol. 108, no. 4, pp. 485–532, 2020.

[5] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv: Computation and Language*, 2020.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[7] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, J. Pino, Alexei Baevski, Alexis Conneau, and Michael Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *Interspeech*, 2021.

[8] Anmol Gulati, James Qin, Chung Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented transformer for speech recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2020-October, pp. 5036–5040, 5 2020.

[9] Yu Zhang, James Qin, Daniel S. Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V. Le, and Yonghui Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," 2022.

[10] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[11] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Y. Qian, Yao Qian, Micheal Zeng, and Furu Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal on Selected Topics in Signal Processing*, 2021.

[12] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[13] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, C. McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," *ArXiv*, 2022.

[14] OpenAI, "Introducing whisper," `https://openai.com/research/whisper`, 2022, Accedido en 2024-03-10.

[15] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, Oct. 2020, pp. 38–45, Association for Computational Linguistics.

[16] OpenAI, "Openai website," `https://openai.com/`, 2024, Accedido en 2024-03-10.

[17] Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," 2021.

[18] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[19] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert, "Mls: A large-scale multilingual dataset for speech research," *ArXiv*, vol. abs/2012.03411, 2020.

[20] Andrew C. Morris, Viktoria Maier, and Phil Green, "From wer and ril to mer and wil: Improved evaluation measures for connected speech recognition," *8th International Conference on Spoken Language Processing, ICSLP 2004*, pp. 2765–2768, 2004.