www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 16(4), Sep-Dec 2025, 364-374. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2025.v16i4.900

Towards a Clinical Interface for Speaker Identification and Speech-To-Text Transcription for Recording Medical Consultations in Spanish

Jonathan Zavala-Díaz[1], *Juan C. Olivares-Rojas*[1], *José A. Gutierrez-Gnecchi*[1], *Adriana C. Tellez-Anguiano*[1], *Enrique Reyes-Archundia*[1], *J. Guadalupe Ramos-Díaz*[1]
[1] División de Estudios de Posgrado e Investigación, Tecnológico Nacional de México I. T. de Morelia, México
{d19123006, juan.or, jose.gg3, adriana.ta, enrique.ra, jose.rd}@morelia.tecnm.mx

Abstract. This paper presents the development of an advanced clinical interface built on the LattePanda Sigma, an embedded device designed for edge computing. The interface integrates OpenAI language models and Whisper for automated speech-to-text transcription, together with accurate speaker diarisation in clinical settings using the pyannote/speaker-diarization-3.1 model. A dataset of ten doctor–patient conversations in Spanish—translated and re-recorded to suit the local context—was used to evaluate the models. Automatic transcriptions generated by the models were compared with the reference transcripts using the ROUGE metric. Average ROUGE scores of 0.9028 for the Small model and 0.9260 for the Medium model indicate high transcription accuracy. The reference transcripts were also used to assess the segments identified by the pyannote model. Finally, the paper analyses the system's usefulness and effectiveness in improving Spanish-language clinical records.
Keywords: NLP, speech-to-text, speaker segmentation, clinical interface

Article Info
Received February 25, 2025
Accepted July 6, 2025

1 Introduction

Electronic medical records (EMRs) have changed how physicians handle patient information, making organizing notes and collecting data easier. However, while they offer clear advantages, they also pose new challenges. Many physicians feel overwhelmed by the paperwork required and spend valuable time entering data instead of caring for their patients. Strict rules make data entry slower than it should be, taking away time that could be better used for patient care. As a result, physicians spend more time than necessary on administrative tasks, reducing the time available for direct patient care and contributing to burnout (Avendano et al., 2022).

The increase in clinical professionals experiencing burnout has been associated with a high administrative burden. To mitigate this problem, technologies such as automatic speech recognition (ASR) and natural language processing (NLP) offer the opportunity to automate clinical documentation using a "digital scribe" (van Buchem et al., 2021).

This paper presents a prototype system that integrates state-of-the-art language models to improve speaker segmentation and speech-to-text conversion within clinical consultations. The resulting tool is embedded in a local hardware device. It offers an intuitive interface for real-time transcription and labeling of medical dialogues, assisting physicians in reducing documentation time and improving continuity of care.

The main contributions of this work are as follows. First, we built a Spanish-language dataset for clinical conversations based on the original English dataset of 272 dialogues by (Fareez et al., 2022). To achieve this, we employed two complementary strategies. On one hand, the whole dataset was automatically translated using the facebook/nllb-200-distilled-600M model and subsequently converted to speech with Azure's neural Spanish voices (es-MX-JorgeNeural and es-MX-DaliaNeural). On the other hand, a subset of ten dialogues was manually recorded by two Spanish-speaking participants simulating doctor-patient consultations, in order to incorporate natural prosody and human voice diversity. This hybrid dataset enables robust testing of diarization and speech-to-text models in both synthetic and real settings.

Second, we performed a systematic evaluation and integration of state-of-the-art speaker diarization and speech recognition models (Pyannote and Whisper) within the context of Spanish clinical dialogues. We analyze performance using ROUGE(Recall-Oriented Understudy for Gisting Evaluation) metrics and manual verification of speaker labels.

Finally, we designed and implemented a working prototype that integrates these models into a virtual assistant intended to support real-time clinical documentation in Spanish. This prototype demonstrates the practical feasibility of such technologies for reducing administrative burden in healthcare settings.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 presents the theoretical framework. Section 4 details the methodology, including consolidation of recording datasets for testing, model evaluation, and system development. Section 5 analyzes the obtained results, while Section 6 discusses findings and limitations. Finally, Section 7 presents the conclusions and future work

2 Related works

The authors in (Avendano et al., 2022) discuss different innovative ways to reduce the burden of data entry into EMRs, including Voice Recognition techniques and Digital Scribes (artificial intelligence). This gives us a perspective on integrating technologies such as NLP and Artificial Intelligence in a clinical environment, specifically for data ingestion into EMRs.

In recent years, interest in digital scribes and conversational assistants in healthcare has grown. Van Buchem et al. (2021) reviewed various systems and found that generic ASR systems exhibit Word Error Rates (WER) up to 65%, while models trained on clinical dialogues reduce WER to around 18%. However, their evaluations rely on English metrics and rarely address real-world usability. Tran et al. (2022) compared generic and specialized ASR for English doctor–patient conversations, reporting WER between 8.8% and 10.5% and word diarization error rates (WDER) from 1.8% to 13.9%. Although these results suggest clinical models can be helpful under ideal conditions, errors persist. In contrast, our work applies pre-trained Whisper (ASR) and Pyannote (speaker diarization) directly to real Spanish consultations, incorporates synthetic data, and validates outcomes using ROUGE metrics and human annotations, while deploying a functional prototype.

Sezgin et al. (2023) developed an emergency digital scribe that uses LLMs (T5, BART, PEGASUS) to summarize doctor–client transcripts in English, achieving ROUGE-1 ≈ 0.49 with fine-tuned BART. In contrast, our study focuses on transcription and speaker diarization in Spanish clinical consultations, using pre-trained models in this language and evaluating with ROUGE metrics.

Fernández Rodríguez (2022) created a Spanish digital scribe using Google Speech-to-Text to transcribe simulated consultation audio, achieving a WER of 9–10% and strong F1 scores in entity extraction. However, their prototype relies on synthetic data and a cloud-based commercial ASR service. In contrast, we employ Whisper and Pyannote—models that can run locally—to process real Spanish clinical consultation recordings, overcoming language barriers and reliance on external services.

Seth et al. 2024 define “ambient scribes” as systems that “listen” to consultations to automatically generate clinical documentation, combining ASR and NLP to produce notes or summaries. While many studies—more conceptual or focused on reducing administrative burden—point out deficiencies in diarization and contextual understanding, our work presents a practical solution in Spanish, evaluating both transcription accuracy and the quality of speaker segmentation.

In general, AI-assisted transcription in medicine (e.g., WER < 11% in English or Spanish) has proven viable, but it is often limited to a single language, domain, or metric. Differing from previous approaches, our work evaluates pre-trained models directly in Spanish using synthetic and real data, integrates them into a functional prototype, and combines automatic transcription with speaker segmentation. We validate results using modern metrics (ROUGE) and human annotations.

3 Theoretical Framework

3.1 Speaker segmentation

Speaker segmentation, also known as speaker diarization, is defined, according to the authors in (Anguera Miro et al., 2012), as the task of determining "who spoke and when" in an audio or video recording that contains an indeterminate amount of speech and an unknown number of speakers.

Speaker diarization was initially introduced as part of ASRresearch, serving as an early step in the process. Over the years, however, this technology has evolved and become a key tool for a variety of applications, including navigation, information retrieval, and advanced analysis of audio data (Anguera Miro et al., 2012).

Diarization systems mostly employ unsupervised machine learning algorithms to analyze utterances exchanged between speakers without knowing in advance the specific labels for each one (Khoma et al., 2023). In this context, Pyannote.audio emerges as a Python library for speaker diarization and audio analysis. It is designed to work with audio data in speaker identification and segmentation, facilitating tasks such as speaker separation and temporal annotation in recordings (Bredin, 2023).

3.2 Speech to Text Conversion

Speech-to-text conversion is the process of transforming speech into written text. Although often confused with speech recognition, the latter term encompasses a more general process of speech understanding (Trivedi et al., 2018). Within this field, Whisper, an ASR system developed by OpenAI, was trained on 680,000 hours of multilingual and multitasking data from the web. This advanced model not only performs accurate transcriptions in multiple languages but can also translate and handle variations in accents and acoustic conditions, thanks to its diverse training. It uses deep neural networks and machine learning techniques to effectively process and convert speech to text, useful in applications such as audio transcription, subtitle generation, and language translation. Its robustness and versatility make it ideal for facing challenges in real-world situations (Radford et al., 2023).

3.3 Rouge Metric

ROUGE is a widely used metric for evaluating automatically generated summaries and texts. Among its variants, ROUGE-1 and ROUGE-L are particularly popular. These metrics focus on assessing the quality of the generated output by comparing it to a reference summary or text.

ROUGE-1 assesses summary quality by comparing the word match between the generated and reference summaries. It is based on unigram n-grams (single words) and evaluates three aspects of performance. Precision measures the proportion of n-grams in the generated summary that are also present in the reference summary. At the same time, recall indicates the proportion of n-grams in the reference summary that appear in the generated summary. Finally, F1-Score, defined as the harmonic mean of precision and recall, provides a balanced metric that captures both dimensions simultaneously.

ROUGE-L measures summary quality by considering the length of the most extended sequences of matching words (word subsequences). It evaluates the length of the longest word subsequence in the generated and the reference summaries, reflecting the text's consistency and fluency. ROUGE-L precision refers to the proportion of matching word subsequences in the generated summary compared to the reference. At the same time, recall indicates the proportion of matching subsequences in the reference summary that appear in the generated one. Finally, the ROUGE-L F1-score, defined as the harmonic mean of precision and recall, provides a balanced measure of performance.

3.4 Hardware Implementation

The Latte Panda Sigma embedded system was used as the primary device for the hardware implementation, and all models developed in this work were executed locally. The device includes an Intel Core i5-1340P processor (12 cores, 3.40–4.60 GHz turbo mode). It also includes 16 GB of LPDDR5 memory at 6400 MHz and a 500 GB solid-state drive, operating under Windows 11. In addition to the central system, a touch screen was employed to visualize the transcripts and segment the speakers in the doctor–patient dialogues. A microphone was used to capture the recordings, while a keyboard and mouse served as complementary input devices.

4 Methodology

The methodology used consists of 6 steps. The first step involves consolidating a test data set that contains doctor-patient dialogues. The second step is evaluating the speaker segmentation model using this data set. In the third step, the speech-to-text conversion models are evaluated. Subsequently, in the fourth step, an analysis of the speech-to-text conversion and segmentation models is performed. Once this analysis is done, the models are implemented in an interface developed and executed on hardware as a final tool. Finally, in the sixth step, the results and our final interface are analyzed and evaluated. Figure 1 illustrates the proposed methodology.

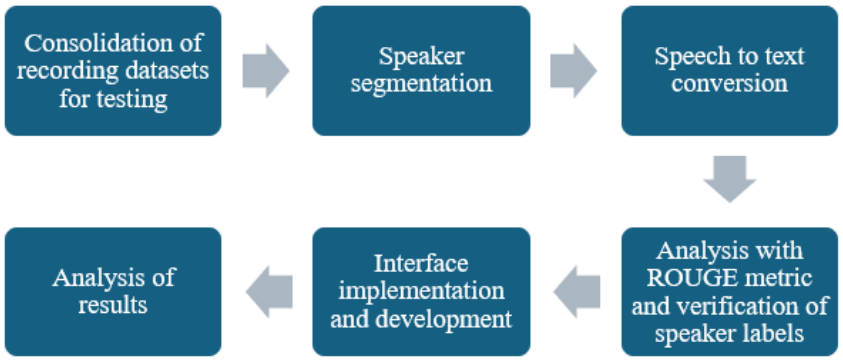


Fig. 1. Methodology.

4.1 Consolidation of recording datasets for testing

The dataset presented in (Fareez et al., 2022) covers a set of medical conversations in Objective Structured Clinical Examinations (OSCE) format, focusing on respiratory cases in audio format and the corresponding text documents. These cases were simulated, recorded, transcribed, and manually corrected to provide a comprehensive dataset of medical conversations to the academic and industrial community.

However, since the dataset is in English and our applications are intended for a Spanish-speaking population, it is necessary to translate the data. Therefore, the first stage of the methodology consists of translating the conversations from English to Spanish. To do this, we took 10 transcripts of English conversations from (Fareez et al., 2022) and translated them into Spanish using Google Translate. Once the translations were completed, we proceeded to the next stage: recording conversations in Spanish. For this, two participants acted as doctors and patients, and the recordings were made on a computer using a USB condenser microphone. The final dataset is generated with the recordings and the transcripts of the 10 clinical notes. Figure 2 shows the sequence of steps for consolidating the set of recordings.

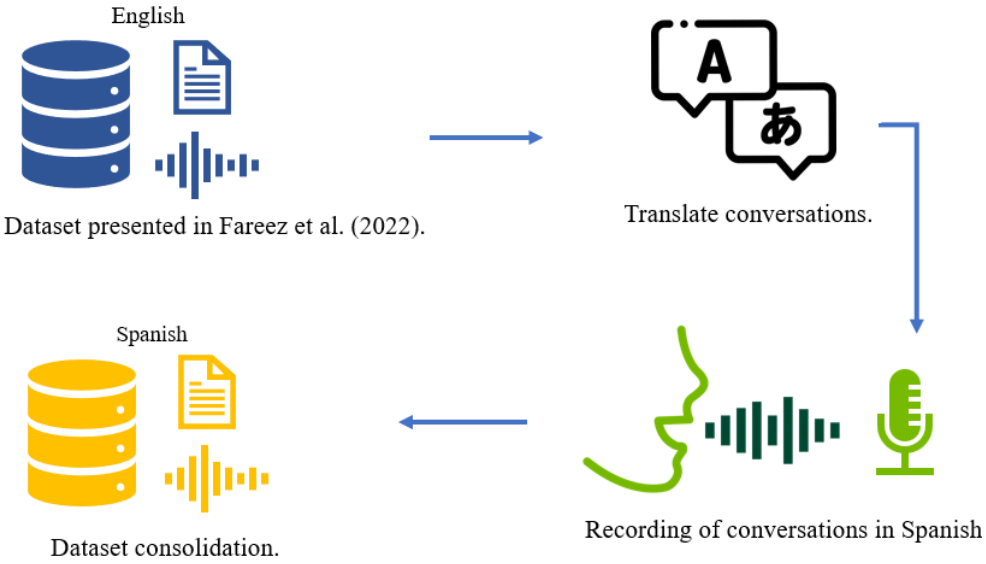


Fig. 2. Consolidation of recording datasets for testing.

To adapt the original English dataset by Fareez et al. (2022) to Spanish, we automatically translated all 272 transcripts using the facebook/nllb-200-distilled-600M model. Then, we generated synthetic speech using Azure's neural voices (es-MX-JorgeNeural and es-MX-DaliaNeural), simulating realistic doctor-patient conversations. This allowed us to run the full system pipeline (speaker diarization with Pyannote and transcription with Whisper) over the entire dataset, enabling large-scale evaluation in Spanish prior to real-world testing.

4.2 Speaker segmentation

For speaker segmentation, the pyannote.speaker-diarization-3 model was used, implemented in Python using the pyannote.audio library. This model processes an audio file in WAV format corresponding to the medical conversation and provides labels for each speaker (e.g., speaker 1, speaker 2, etc.) and the start and end times at which each speaker is present in the recording.

4.3 Speech-to-text conversion

The OpenAI Whisper model, implemented in Python using the Hugging Face library, was used in its small and medium versions for the speech-to-text conversion process. The process begins with the reception of an audio file in WAV format containing the conversation between doctor and patient. The Whisper model analyzes this audio file and generates a written conversation transcript.

4.4 Analysis with ROUGE metric and verification of speaker labels

For the ROUGE metric analysis, we used the Python rouge-score library, which generates ROUGE-1 and ROUGE-L metrics, including precision, recall, and F1-score. This analysis compares the original transcript to the one generated by the OpenAI Whisper models. Speaker 0 and Speaker 1 are first identified for manual verification of speaker labels. Once identified, the labels are changed to "doctor" or "patient". If two consecutive lines belong to the same speaker, they are merged into one. Subsequently, the labels are manually compared to the original transcript to see if they are correct, and a count of the incorrect labels is made.

4.5 Interface implementation and development

For the development and implementation of the interface, as shown in Figure 3, the recording of the doctor-patient dialogue during the medical consultation is schematized using a microphone. This recording is processed locally on the Latte Panda Sigma embedded device, where the speaker segmentation and speech-to-text conversion models are executed. These models are implemented in the Python programming language. It is also used to develop an intuitive graphical interface with buttons to start or stop the recording and perform its processing. At the end, the transcription of the doctor-patient dialogue is obtained.

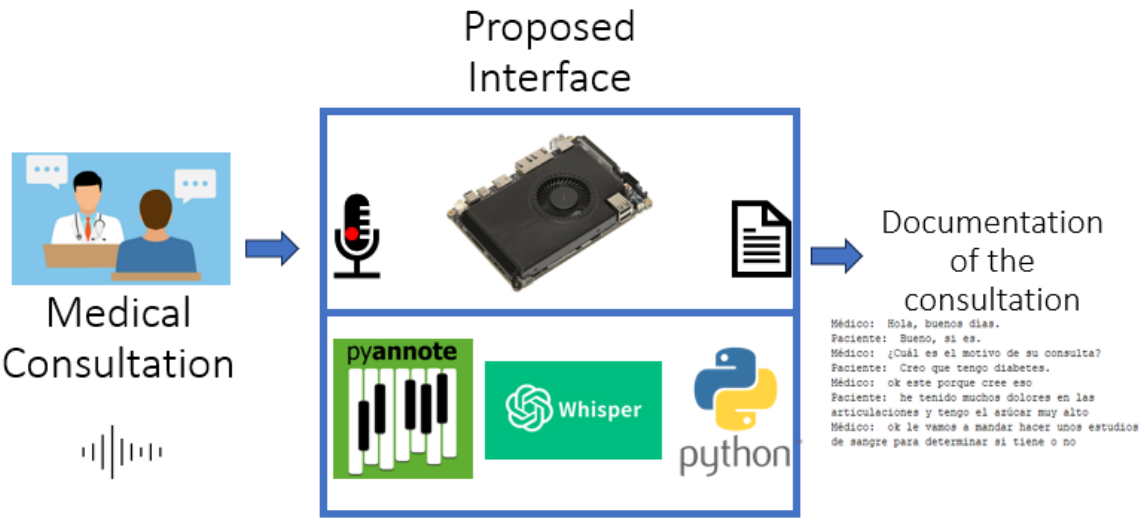


Fig. 3. Schematic of the proposed interface

4.6 Prototype testing

The results obtained from the prototype are preliminarily examined, with a detailed discussion presented in Section 5. This stage evaluates the feasibility of implementing the proposed tools in real medical practice to streamline clinical documentation during patient consultations.

5 Analysis of results

A dataset specific to clinical conversations was created, starting from the English transcripts of (Fareez et al., 2022). The process included two key stages: first, the original English transcripts were translated into Spanish, and second, new Spanish conversations were recorded based on the translated transcripts. The resulting dataset includes 10 Spanish recordings, in WAV format, corresponding to interactions between doctors and patients in a medical consultation setting. Each recording is accompanied by its Spanish transcript, which is carefully labeled with the identities of "patient" and "doctor" for each dialogue. This dataset supports the evaluation and development of NLP applications in Spanish-speaking clinical settings. It includes both audio and annotated transcripts, enabling detailed analysis.

In addition, the same 10 dialogues were selected from the complete set of 272 conversations from equivalent synthetic versions. These versions are produced by translating the original transcripts into Spanish and converting them into audio using neural voices. This resulted in a parallel subset containing both real recordings—produced by humans in a controlled setting—and synthetically generated versions from the duplicate textual content. This dual design allows for a direct comparison between human speech and synthesized speech, facilitating a more robust evaluation of the performance of the segmentation and transcription models in Spanish-language clinical contexts.

Once the dataset preparation was completed, we proceeded to deploy the pyannote.speaker-diarization-3 model for speaker identification. This model receives audio files in WAV format corresponding to the medical conversations in our Spanish dataset and provides speaker identifications along with each intervention's start and end times.

With this information, we cut the original audio into segments based on the identified time intervals, assigning the corresponding labels to each speaker. For example, for an audio file, the results can be: Speaker_00: start 00:01 – end 00:05, and Speaker_01: start 00:06 – end 00:10.

Once the segmentation is done, we transcribe each audio segment into text using OpenAI's Whisper model. This model converts audio segments into written text, allowing for detailed analysis of medical conversations. Two versions of the Whisper model were implemented for this process: small and medium. The choice of these versions is based on the need to evaluate the effectiveness of the models based on transcription accuracy and overall performance. The small version is faster and uses fewer resources, while the medium version is more accurate but demands more computing power. Comparing them helps decide which works best based on audio quality, speech clarity, and medical vocabulary. This dual approach ensures a thorough evaluation of the model's capabilities under different conditions and helps to select the most suitable tool for future applications in the clinical setting.

The process described was applied to all 10 recordings in our dataset. Figure 4 illustrates a segmented and transcribed recording using Pyannote.speaker-diarization-3 model for speaker identification and Whisper Medium for speech-to-text conversion. In Figure 4a, the labels assigned to the speakers are Speaker_01 and Speaker_00, which correspond to D: Doctor and P: Patient, respectively. Since the speaker identification process is not automated, the correspondence between the labels Speaker_01 and Speaker_00 and the roles of Doctor and Patient was performed manually, as observed in Figure 4b. This manual adjustment ensures a correct assignment of the labels and an accurate interpretation of the transcripts in the medical context. The visualization in the figures allows us to evaluate the accuracy and usefulness of the segmentation and transcription, facilitating the analysis of medical conversations.

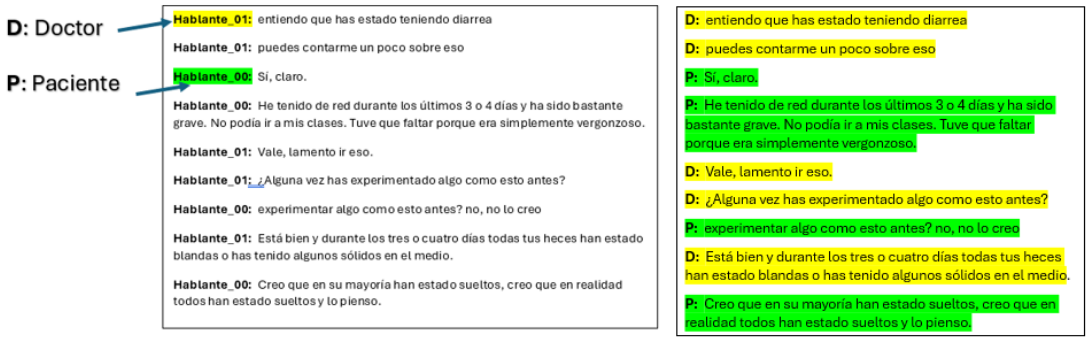


Fig. 4. Transcription, a) Identification, b) Changing labels.

Once the labels have been changed, in Figure 5b, we can see consecutive segments labeled towards the same speaker, for example, label D. *D: "entiendo que has estado teniendo diarrea" and D: "puedes contarme un poco sobre eso"*. This difference is because during the recording, there may be silences or pauses that the speaker segmentation model identifies as separate segments. However, the label "D" is correctly assigned to all interventions by the same speaker. To make our transcriptions of doctor-patient dialogues more fluid, we merge the consecutive interventions of the speakers into one.

Figure 5a shows the transcript obtained by joining consecutive interventions of the speakers. This was then manually compared with the original transcript of our test data set, which is shown in Figure 7b. For this comparison example, we notice that the labels are placed correctly by looking at the transcript obtained with the implemented models in Figure 5a and the original transcript from our database in Figure 5b.

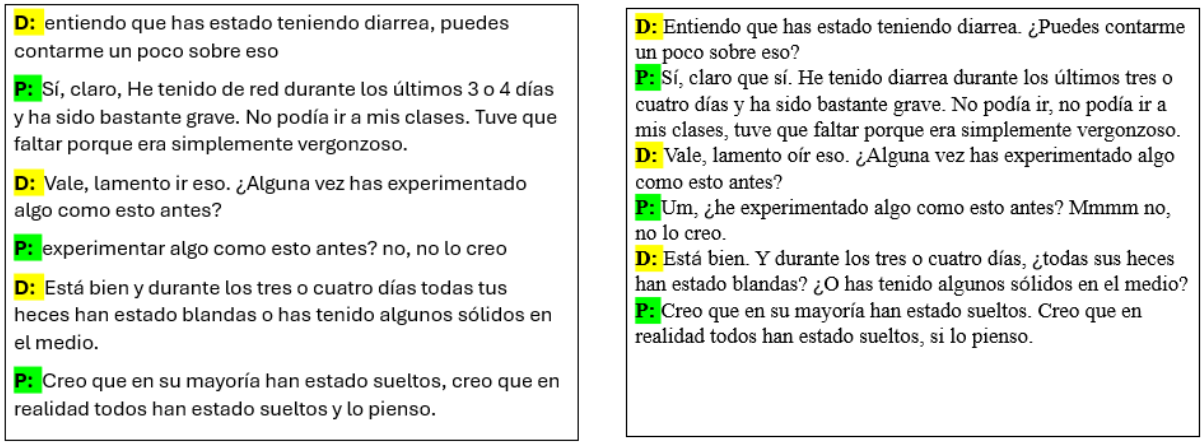


Fig. 5. Transcription, a) Implemented, b) Original.

Table 1 presents the manual evaluation of speaker labels for each audio file, comparing the performance of the diarization system on both real and synthetic recordings of the same conversations. For the real recordings, an average precision of 0.98 was achieved, with four conversations (CAR0001, CAR0002, GAS0002, and GAS0003) labeled perfectly (100% accuracy). The highest number of incorrect labels occurred in GAS0001, with five mislabeled segments, mainly corresponding to short utterances such as "No." In contrast, the synthetic versions showed a lower average precision of 0.84. The most significant discrepancy appeared in conversation GAS0001, where the system misclassified 28 out of 77 segments (precision of 0.67). These results suggest that although the diarization model performs robustly on real speech, synthetic voices can introduce greater variability in acoustic patterns, affecting the accuracy of speaker segmentation. Nonetheless, the overall label placement is essentially correct across both types of input.

Table 1. Manual evaluation of labels

Conversation	Label (Recording)			Precision	Label (Synthetic)		Precision
	Total	Correct	Incorrect		Correct	Incorrect	
CAR0001	122	122	0	1	90	32	0.74
CAR0002	83	83	0	1	73	10	0.88
CAR0003	123	122	1	0.99	109	14	0.89
CAR0004	69	66	3	0.95	61	8	0.88
CAR0005	69	67	2	0.97	63	6	0.96
DER0001	86	85	1	0.98	77	9	0.90
GAS0001	77	72	5	0.94	49	28	0.67
GAS0002	79	79	0	1	68	11	0.86
GAS0003	87	87	0	1	70	17	0.80
GAS0004	81	79	2	0.98	70	11	0.86
			Average	0.98		Average	0.84

To evaluate the quality of the speech-to-text conversion, we used the ROUGE metric to compare the similarity between the original text and the generated transcripts. The implementation of this metric was done in Python using the rouge_scorer library. This evaluation included comparing the results obtained with the Whisper small and Whisper medium versions. The ROUGE metric measures the quality of the transcripts based on the match of n-grams, word sequences, and other linguistic units between the generated text and the original text. By analyzing these metrics, we can determine how faithful each version of the model is to the information contained in the original audio and choose the version that offers the highest accuracy and consistency in the speech-to-text conversion.

Table 2 presents the ROUGE metric results for audio transcription using Whisper small and medium models on both real and synthetic audio. The values correspond to the average of 10 evaluated transcripts and include precision, recall, and F1-score for ROUGE-1 and ROUGE-L.

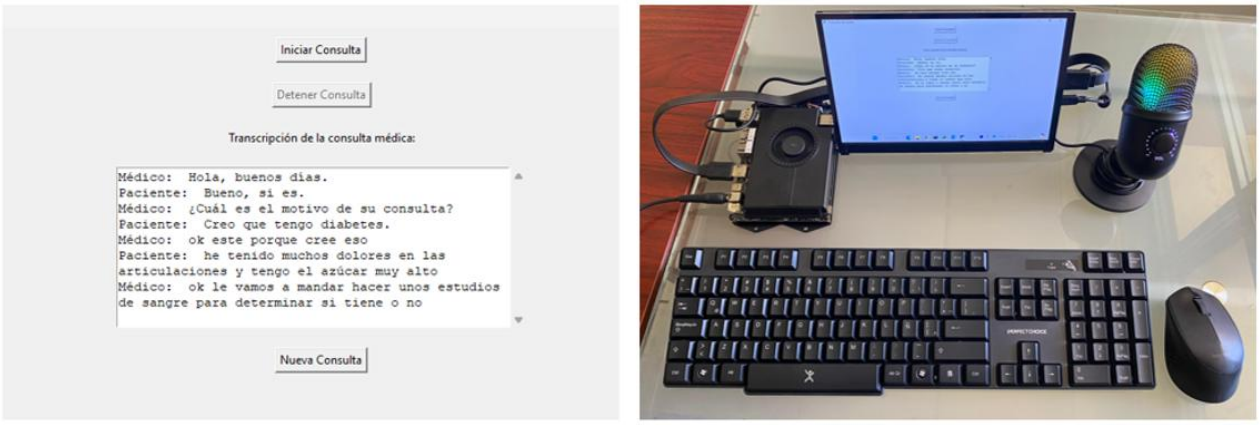
The Whisper medium model outperforms Whisper small across both metrics for real recordings, achieving the highest scores with a ROUGE-1 F1 of 0.9335 and a ROUGE-L F1 of 0.9219. This indicates a better ability to preserve word content and sequence accuracy during transcription. In contrast, performance decreases when applied to synthetic audio: although Whisper small maintains relatively high ROUGE-1 precision (0.8659), recall drops significantly (0.7344), with an F1-score of 0.7938. Similar trends are observed for ROUGE-L.

These results suggest that Whisper medium is better suited for high-quality transcription tasks, especially with real human speech. The lower performance on synthetic speech highlights the need for further model adaptation or preprocessing when working with generated audio in clinical NLP applications.

Table 2. ROUGE Metric Results

		Recording		Synthetic
Model	Metric	Small	Medium	Small
ROUGE-1	Precision	0.9028	0.9260	0.8659
	Recall	0.9197	0.9413	0.7344
	F1 Score	0.9111	0.9335	0.7938
ROUGE-L	Precision	0.8846	0.9145	0.7476
	Recall	0.9012	0.9297	0.6348
	F1 Score	0.8928	0.9219	0.6858

Figure 6 shows the interface for speaker segmentation and speech-to-text transcription implemented as a final tool. Figure 6a shows the main screen of the developed interface, where three classic buttons can be seen: start query, stop query, and new query. A text box is also shown at the end of the query, where the dialogue is displayed with the speaker segmentation labels and the transcribed text of the dialogue. Figure 6b shows the interface implemented as a final tool in our embedded device, with a microphone for audio capture and a touch screen for displaying information and manipulating the interface. In addition, a mouse and keyboard were used.



a) b)

Fig. 6. User interface, a) Screenshot interface, b) Hardware interface

Figure 7 presents results from five real recordings of simulated clinical interviews conducted with the support of a family physician. Each row shows the original duration of the recording (in seconds) and the corresponding processing times for speaker segmentation (using Pyannote) and transcription (using two Whisper variants: "small" and "medium"). All components were executed locally on our embedded LattePanda interface. As shown, the total processing time increases with audio duration and model complexity but remains feasible for real-time or near-real-time deployment in clinical settings.

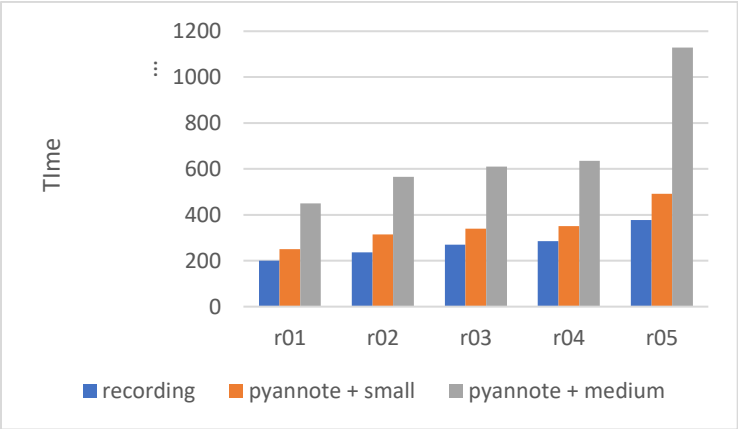


Fig. 7. Processing Time for Local Transcription and Diarization of Clinical Recordings"

6 Discussion

Regarding the dataset used, all 272 conversations from the original English dataset (Fareez et al., 2022) were automatically translated into Spanish and synthesized using neural voices to generate a complete audio corpus. Additionally, 10 of these conversations were recorded with authentic human voices to enable direct comparison between synthetic and real speech. While this synthetic dataset provides a scalable and consistent resource for clinical NLP tasks, including more real voice recordings would enhance the evaluation of model performance under natural acoustic and conversational conditions.

As future work, we propose expanding the number of real recordings and conducting more evaluations in actual clinical environments. This would strengthen the study's findings, improve the generalizability of the results, and ensure that the models perform robustly in realistic settings. A more diverse and representative dataset—both in terms of speakers and clinical scenarios—will offer a solid foundation for future research in speech-to-text transcription and speaker segmentation within Spanish-speaking clinical contexts.

The pyannote.speaker-diarization-3 model showed positive results in the speaker segmentation task, labeling them as Speaker_01 or Speaker_00, respectively, for Doctor and Patient. Although the model performed accurate segmentation, the automatic assignment of the labels Speaker_01 and Speaker_00 to the roles of Doctor and Patient still requires manual intervention.

To make progress in this area, it is proposed to develop an automated system that identifies which of these labels corresponds to the Doctor and which to the Patient. This automatic identification can be approached in several ways. One option is to detect keywords in the conversation that clearly distinguish between the Doctor and the Patient. For example, doctors typically use medical terms or specific phrases, while patients are more likely to ask questions or give answers.

Another strategy could be establishing a protocol in which the first speaker in each conversation is always the Doctor or the Patient. This approach would help to standardize the assignment of labels and reduce the need for manual intervention.
The current system consistently identified interlocutors without mixing the roles of Doctor and Patient. However, automating this task would be a significant advance, improving efficiency and accuracy in the processing of clinical conversations.

Regarding the accuracy of speech-to-text transcriptions, both the small and medium versions of OpenAI Whisper models provided very satisfactory results. The medium version performed exceptionally well, reaching 0.9260 in ROUGE-1 and 0.9145 in ROUGE-L. These scores show a substantial similarity to the original transcription, demonstrating the model's accuracy in capturing audio content.

Other audio preprocessing techniques could further enhance transcription quality. This might involve reducing noise and interference, which can affect accuracy. Applying noise removal filters, volume normalization, and speech clarity enhancement could help assess whether audio quality significantly affects the final transcription. This additional approach could provide more complete insight into optimizing model results and improving accuracy in clinical settings.

7 Conclusions

NLP has proven to be a versatile and powerful tool in the clinical setting. Its applications range from optimizing data management to improving medical decision-making. As we continue to explore new applications and overcome technical and ethical challenges, the potential for NLP to transform healthcare and biomedical research remains promising.

The fusion of these two models can generate accurate transcripts of medical consultations, correctly identifying the speakers. Implementing them in an embedded system as a clinic virtual assistant would facilitate monitoring and recording the patient's clinical history.

The pyannote.speaker-diarization-3 model excelled in speaker segmentation, accurately identifying the interlocutors in consultation recordings, with an average accuracy of 0.98 labelings. Meanwhile, the OpenAI Whisper model effectively converted speech to text, offering accurate and consistent transcripts and obtaining a ROUGE metric accuracy score of 0.9260 for the Medium model.

Bringing these technologies into clinical settings helps streamline consultation documentation and makes it easier to integrate with electronic records. This, in turn, allows for better data extraction and analysis. The results show how artificial intelligence can improve the management of clinical information, opening the door to more efficient and accurate tools for healthcare.

8 Acknowledgment

The authors thank the Tecnológico Nacional de México (TecNM) for the financial support provided in project 21636.25-P. We also thank the Delfin Program for the support provided by the student Carolain Jimenez Bedoya from the Universidad del Valle, Colombia, for their support in the Dataset construction and testing. Finally, Jonathan Zavala-Díaz thanks SeCiHTI for the 2023-000002-01NACF-01652 doctoral scholarship.

References

Anguera, M. X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(2), 356–370. https://doi.org/10.1109/TASL.2011.2125954

Avendano, J. P., Gallagher, D. O., Hawes, J. D., Boyle, J., Glasser, L., Aryee, J., & Katt, B. M. (2022). Interfacing With the Electronic Health Record (EHR): A Comparative Review of Modes of Documentation. *Cureus*. https://doi.org/10.7759/cureus.26330

Bredin, H. (2023). pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. *INTERSPEECH 2023*, 1983–1987. https://doi.org/10.21437/Interspeech.2023-105

Fareez, F., Parikh, T., Wavell, C., Shahab, S., Chevalier, M., Good, S., De Blasi, I., Rhouma, R., McMahon, C., Lam, J.-P., Lo, T., & Smith, C. W. (2022). A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, *9*(1), 313. https://doi.org/10.1038/s41597-022-01423-1

Fernández Rodríguez, M. A. (2022). Transcripción y extracción automáticas de información clave desde audios clínicos en español. https://repositorio.uchile.cl/handle/2250/187613

Khoma, V., Khoma, Y., Brydinskyi, V., & Konovalov, A. (2023). Development of Supervised Speaker Diarization System Based on the PyAnnote Audio Processing Library. *Sensors*, *23*(4), 2082. https://doi.org/10.3390/s23042082

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of Machine Learning Research*, *202*, 28492–28518.

Seth, P., Carretas, R., & Rudzicz, F. (2024). The Utility and Implications of Ambient Scribes in Primary Care. *JMIR AI*, *3*, e57673. https://doi.org/10.2196/57673

Sezgin, E., Sirrianni, J., & Kranz, K. (2023). *Development and Evaluation of a Digital Scribe: Conversation Summarization Pipeline for Emergency Department Counseling Sessions towards Reducing Documentation Burden*. https://doi.org/10.1101/2023.12.06.23299573

Tran, B. D., Mangu, R., Tai-Seale, M., Lafata, J. E., & Zheng, K. (2022). Automatic speech recognition performance for digital scribes: a performance comparison between general-purpose and specialized models tuned for patient-clinician conversations. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2022*, 1072–1080. http://www.ncbi.nlm.nih.gov/pubmed/37128439

Trivedi, A., Pant, N., Shah, P., Sonik, S., & Agrawal, S. (2018). Speech to text and text to speech recognition systems-A review. *IOSR Journal of Computer Engineering (IOSR-JCE)*, *20*(2), 36–43. https://doi.org/10.9790/0661-2002013643

van Buchem, M. M., Boosman, H., Bauer, M. P., Kant, I. M. J., Cammel, S. A., & Steyerberg, E. W. (2021). The digital scribe in clinical practice: a scoping review and research agenda. *Npj Digital Medicine*, *4*(1), 57. https://doi.org/10.1038/s41746-021-00432-5