Healthcare of the Future 2025
T. Bürkle et al. (Eds.)

doi:10.3233/SHTI250213

AI-Enhanced Speech Recognition in Triage

Ahmed ELHILALI[a*], Vanessa BRÜGGER[a*] , Isabelle TSCHANNEN[b], Wolf HAUTZ[b], Gert KRUMMREY[a,b1]
[a] *Bern University of Applied Sciences, Institute for Medical Informatics*
[b] *Inselspital, Bern University Hospital, Department of Emergency Medicine*

ORCiD ID: Wolf Hautz https://orcid.org/0000-0002-2445-984X ,
Gert Krummrey https://orcid.org/0000-0002-8397-2336

Abstract. Triage is used in emergency departments to ensure timely patient care according to urgency of treatment. However, triage accuracy and efficiency remain challenging due to time-constraints and high demand. This proof-of-concept study evaluates an AI-powered triage system that leverages speech recognition (STT) and large language models (LLMs) to process patient interactions in triage and to assign an Emergency Severity Index (ESI) triage level and a classification of the main presenting complaint according to the Canadian Emergency Department Information System (CEDIS). In Switzerland, different Swiss German dialects add to the complexity of the task. STT models achieved word error rates (WER) of 2.3% for High German and 17.66% for Swiss German. Despite the high WER, the AI's classification accuracy reached 90–100% for ESI levels and CEDIS codes. These results highlight the potential of integrating AI into triage workflows, enhancing consistency and reducing the documentation burden for clinical staff. Future research should address multi-language adaptation and data security to ensure seamless implementation in real-world settings.

Keywords. Artificial Intelligence, Speech-to-Text, Triage Systems, Natural Language Processing (NLP), Emergency Medicine, Triage

1. Introduction

Emergency departments (EDs) are pivotal in delivering timely and life-saving care. Over the past decades, the demand for ED services has surged, placing immense pressure on resources and staff. Triage is the initial patient assessment process which determines the urgency of treatment based on the severity of a patient's condition and prioritizes resources accordingly [1].

Among the various triage tools, the Emergency Severity Index (ESI) is a widely adopted, validated system used internationally. Despite its prevalence, inconsistencies in its application persist, as studies show only 59.6% accuracy in standardized triage cases, even among experienced nursing staff [2]. Further, in most EDs, a presenting complaint is documented in addition to the urgency of treatment. The Canadian Emergency Department Information System (CEDIS) is widely used for this purpose [3].

Advancements in artificial intelligence (AI), especially in machine learning (ML) and natural language processing (NLP), offer promising avenues to improve triage accuracy and efficiency [4]. While most existing research focuses on retrospective

[1] Corresponding author: Gert Krummrey, Bern University of Applied Sciences, Quellgasse 21, CH-2502 Biel/Bienne, Switzerland. E-Mail: gert.krummrey@bfh.ch
* Both authors contributed equally and share first authorship

analyses of unstructured text, the integration of speech recognition in real-time clinical workflows remains underexplored [5]. AI-powered systems can process audio recordings of patient-provider interactions, convert them into text using speech-to-text algorithms, and provide automated categorizations for ESI and CEDIS classifications [6].

This study investigates the potential of an AI-enhanced triage system that integrates speech recognition and LLMs into triage. By evaluating transcription accuracy and classification reliability, it aims to demonstrate the feasibility of implementing such systems to optimize ED workflows and to reduce the documentation burden on clinical staff.

2. Methods

The AI-enhanced triage system integrates speech-to-text (STT) technology with large language models (LLMs). It processes audio recordings of patient-provider interactions, transcribes the conversations, and determines the ESI triage level and the CEDIS code. It includes mechanisms for data preprocessing, classification, and manual correction to improve usability and accuracy.

The following STT models were utilized for transcription: 1) Wit.ai: A speech recognition platform, optimized for general use cases, but with limitations in domain-specific medical terminology and dialect, 2) Whisper by OpenAI in the "base" configuration and in a Swiss German fine-tuned version: "nizarmichaud/whisper-large-v3-turbo-swissgerman". The system employs LLMs to process transcribed text and produce structured outputs with ESI triage level determination and CEDIS classification. The model "Claude-3-Opus-20240229" from Anthropic was selected for its ability to generate domain-specific analyses and structured outputs while ensuring contextually appropriate responses.

Ten written triage scenarios were created using common presenting complaints. ChatGPT-4o was used to create dialogs based on these scenarios. An audio dataset of triage conversations was then created in both High German and Swiss German using a bespoke audio dialog creation engine utilizing OpenAI's text-to-speech (TTS) capabilities and a Swiss German TTS API [7]. Reverberation and ambient noise were added to the recordings to make them more realistic. Each scenario included annotated ESI levels and CEDIS codes determined by clinical experts to be referenced later in the evaluation.

Using web-technologies and Python for the backend, a web-application was programmed to allow the upload of audio data into the processing pipeline (see Fig. 1). Sequentially, audio was transcribed and the result passed on to the LLM which in turn determined ESI level and CEDIS code as well as providing a summary of the conversation for documentation purposes.

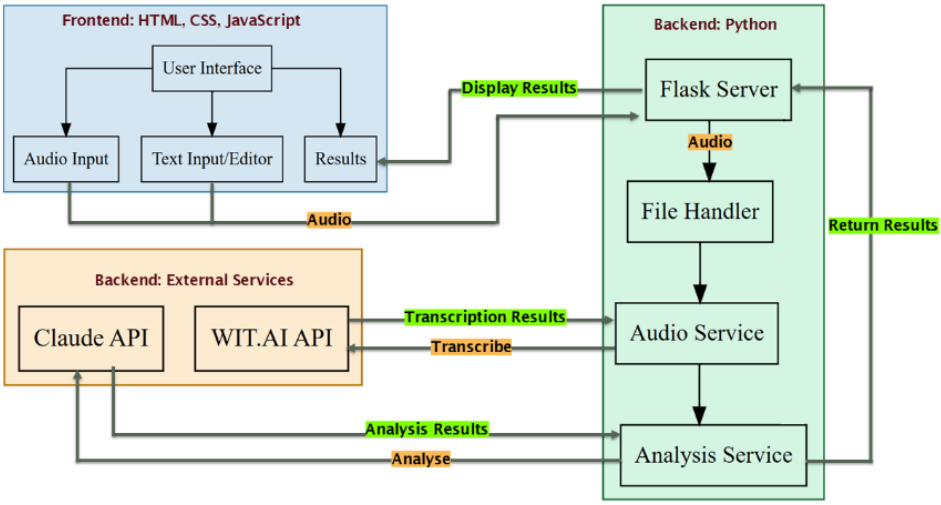

Figure 1. System Design

3. Results

Speech-to-text (STT) models demonstrated significant variations in Word Error Rates (WER) depending on the language and model used. The transcription performance varied across the triage scenarios.

High German: Wit.ai achieved a mean WER of 2.30%, significantly outperforming the base Whisper model, which had a WER of 4.11%.

Swiss German: The adapted Whisper model delivered a mean WER of 17.66% as compared to the base Whisper model (mean WER 45.19%), showcasing a substantial improvement of fine-tuned models over general-purpose models. Wit.ai struggled with Swiss German dialectal variations (mean WER of 63.24%).

Table 1. Word Error Rates (WER) for Swiss German / High German audio

	Swiss German			High German	
Scenario	**Whisper Base**	**Whisper Swiss**	**Wit.ai**	**Whisper Base**	**Wit.ai**
Chest Pain	50.00%	21.30%	67.82%	3.25 %	0.90%
Ear Pain	46.26%	21.90%	62.20%	5.10 %	1.54%
Abdominal Pain	42.41%	18.85%	58.93%	3.66 %	2.29%
Diarrhea	45.85%	17.15%	59.77%	5.42 %	2.18%
Flank Pain	46.82%	14.64%	65.44%	3.28 %	2.74%
Depression	34.30%	18.11%	55.66%	3.28 %	3.61%
Dizziness	49.25%	12.04%	63.45%	2.80 %	2.32%
Headache	50.00%	15.99%	70.91%	2.43 %	2.18%
Back Pain	48.18%	19.06%	68.33%	6.64 %	2.41%
Laceration	38.82%	17.51%	59.88%	5.27 %	2.83%

Classification Accuracy: The system's classification accuracy for ESI level and CEDIS code was evaluated across both transcription formats and input languages:

High German: 90% accuracy for ESI levels and 100% accuracy for CEDIS codes.

Swiss German: 100% accuracy for ESI levels and 90% accuracy for CEDIS codes.

4. Discussion

This proof-of-concept demonstrates the potential of AI-powered triage systems to enhance emergency department (ED) workflows by integrating advanced speech recognition and classification capabilities. ESI code assignments were validated by an experienced ED triage nurse, with synthetic audios incorporating background noise for enhanced authenticity. However, we are aware of the limitations of using synthetic data for triage conversations, which was necessitated by data privacy considerations and the lack of ethics approval for using real patient data.

The Whisper model fine-tuned for Swiss German significantly outperformed general-purpose alternatives like the Whisper base model and Wit.ai for Swiss German transcription, significantly reducing the Word Error Rate. However, in dialects without an accepted way of spelling words, calculating the WER is challenging. Also, all STT-systems fall short of the ideal accuracy for clinical applications, where errors in transcription can propagate through subsequent classification processes. Challenges in processing dialectical variations underline the need for further fine-tuning of STT models tailored to regional languages.

Despite the limitations in transcribing audio, especially when dialect is spoken, the system achieved high accuracy in Emergency Severity Index (ESI) and Canadian Emergency Department Information System (CEDIS) classification, with an accuracy of up to 100% in certain scenarios. This illustrates that for those tasks an understanding of the context seems sufficient for the model to determine the right level and code.

While promising, the system faces significant barriers regarding data security and ethical considerations. Integration with external services like Wit.ai raises compliance concerns with Switzerland's data protection laws. Local hosting and potential bias mitigation are necessary.

This study demonstrates the feasibility of integrating AI-enhanced speech recognition and classification systems into ED triage workflows. With classification reliability reaching acceptable levels for clinical application, the system offers significant potential to reduce the administrative burden on healthcare. Future work should focus on refining dialect-specific STT models to improve transcription accuracy, enabling multi-label classification, and ensuring data privacy through locally hosted solutions.

References

[1] Dugas AF, Kirsch TD, Toerper M, Korley F, Yenokyan G, France D, u. a. An Electronic Emergency Triage System to Improve Patient Distribution by Critical Outcomes. J Emerg Med. Juni 2016;50(6):910–8. doi: 10.1016/j.jemermed.2016.02.026.

[2] Grossmann FF, Delport K, Keller DI. Emergency Severity Index. Notf Rettungsmedizin. 1. Juni 2009;12(4):290–2.

[3] Grafstein E, Unger B, Bullard M, Innes G, Group the CEDIS (CEDIS) W. Canadian Emergency Department Information System (CEDIS) Presenting Complaint List (Version 1.0). Can J Emerg Med. Januar 2003;5(1):27–34. doi: 10.1017/s1481803500008071.

[4] Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. Drazen JM, Kohane IS, Leong TY, Herausgeber. N Engl J Med. 30. März 2023;388(13):1233–9. doi: 10.1056/NEJMsr2214184.

[5] Kwon J myoung, Lee Y, Lee Y, Lee S, Park H, Park J. Validation of deep-learning-based triage and acuity score using a large national dataset. PLoS ONE. 15. Oktober 2018;13(10):e0205836.

[6] McHugh M, Tanabe P, McClelland M, Khare RK. More patients are triaged using the Emergency Severity Index than any other triage acuity system in the United States. Acad Emerg Med Off J Soc Acad Emerg Med. Januar 2012;19(1):106–9. doi: 10.1111/j.1553-2712.2011.01240.x.

[7] Cieliebak, Mark. ZHAW Zürcher Hochschule für Angewandte Wissenschaften. [zitiert 24. Januar 2025]. End-to-End Low-Resource Speech Translation for Swiss German Dialects. Verfügbar unter: https://www.zhaw.ch/de/forschung/projekt/72998