Enhancing Mental Health Support through Artificial Intelligence: Advances in Speech and Text Analysis within Online Therapy Platforms

Mariem Jelassi , Khouloud Matteli , Houssem Ben Khalfallah , Jacques Demongeot *

Posted Date: 27 February 2024

doi: 10.20944/preprints202402.1585.v1

Keywords: Conversational AI; Automatic Speech Recognition (ASR); Natural Language Processing (NLP); Online Therapy Platforms; AI in Mental Healthcare

Article

Enhancing Mental Health Support through Artificial Intelligence: Advances in Speech and Text Analysis within Online Therapy Platforms

Mariem Jelassi [1], Khouloud Matteli [2], Houssem Ben Khalfallah [1,3] and Jacques Demongeot [3,*]

[1] RIADI Laboratory, ENSI, Manouba University, 2010, La Manouba, Tunisia
[2] ESEN, Manouba University, 2010, La Manouba, Tunisia
[3] AGEIS laboratory, UGA, 38700, La Tronche, France
* Correspondence: jacques.demongeot@univ-grenoble-alpes.fr

Abstract: In the dynamic field of mental health care, the nuanced application of Artificial Intelligence (AI) through Natural Language Processing (NLP) and Automatic Speech Recognition (ASR) is a pivotal innovation in patient empowerment and service optimization. This study introduces a distinctive online therapy platform that capitalizes on the synergy of NLP and ASR to offer unprecedented levels of interactive and personalized therapeutic interventions. The architecture of our system is meticulously detailed, featuring an ASR component with an impressive Word Error Rate (WER) of 14% when trained on the diverse French subsets of the Mozilla Common Voice dataset, complemented by a high-precision NLP framework skilled in processing and responding to user inputs. The evaluation of our system highlights its efficacy in enhancing therapy sessions and user satisfaction, with an emphasis on the qualitative aspects of user feedback. The paper addresses challenges such as dataset representativeness and language model refinement and articulates the strategic solutions employed to overcome them. The paper concludes with forward-looking perspectives on AI's role in mental health services, advocating for the creation of sophisticated, language-specific datasets and models to satisfy the increasing demands of online therapy, reflecting a growing commitment of patients in the management of their therapy. This research underscores the transformative impact of AI in advancing mental health care into the digital age, representing a significant evolution over existing methodologies.

Keywords: Conversational AI; Automatic Speech Recognition (ASR); Natural Language Processing (NLP); Online Therapy Platforms; AI in Mental Healthcare

1. Introduction

Digital age has ushered in a plethora of technological advancements, with Automatic Speech Recognition (ASR) standing as a paramount testament to this evolution. From its nascent stages in the mid-20th century, where rudimentary systems could recognize a mere handful of words, ASR has burgeoned into a sophisticated tool capable of transcribing multiple languages, dialects, and accents with a remarkable precision [1]. Such advancements are not mere technological marvels; they hold profound implications for sectors like healthcare, where communication is primordial .

Conversational agents, underpinned by ASR and enriched with Natural Language Processing (NLP) capabilities, are redefining our digital interactions. These agents, extending beyond the realms of mere chatbots, simulate human-like interactions, offering potential applications ranging from e-commerce customer support to intricate medical guidance [4]. As the global demand for efficient and accessible healthcare solutions intensifies, the integration of ASR and NLP in medical applications emerges as a beacon of innovation. This research delves into the transformative potential of these technologies in healthcare, particularly emphasizing their role in enhancing patient-provider communication and psychological well-being.

The medical landscape is witnessing a paradigm shift, with ASR and conversational agents at its epicentre. In oncology, these technologies have transitioned from experimental tools to essential

 doi:10.20944/preprints202402.1585.v1

components, offering patients comprehensive guidance on treatments, potential side effects, and post-treatment recovery [5]. For the elderly, whose increasing demography is often challenged by rapid technological advancements, ASR-integrated devices have become indispensable, monitoring daily activities and ensuring timely medical interventions [6].

Chronic conditions, such as diabetes and hypertension, necessitate rigorous monitoring. Here, ASR and conversational agents offer holistic solutions, encompassing medication reminders, dietary advice, exercise guidelines, and real-time vital variables monitoring [7]. Their transformative impact extends to rehabilitation, aiding patients with speech and mobility impairments [8], and to paediatric care, where they simplify medical terminologies for young patients [9]. In cardiology, the integration of ASR into monitoring devices has paved the way for real-time feedback mechanisms, a leap that holds life-saving potential [10].

The confluence of technology and mental health has birthed a novel approach to psychological and psychiatric care. Conversational agents, fortified with NLP and ASR, are revolutionizing therapeutic interventions. Digital therapists, such as Woebot® (a relational agent for mental health [11]), employ principles from cognitive-behavioural therapy (CBT) to engage users, showcasing significant efficacy in mitigating symptoms of depression and anxiety [12]. Numerous studies underscore the potential of these agents in delivering psychological interventions, with some rivalling the effectiveness of human therapists [13].

In the broader realm of psychiatric care, the applications of these agents are multifaceted. They assist in diagnostic assessments, monitor medication adherence, and provide therapeutic interventions for intricate psychiatric disorders [14]. The integration of ASR enhances their capabilities, enabling real-time vocal interactions that can discern user emotions, sentiments, and potential distress signals [15] combined with digital tools looking for specific behaviours of a pathological condition (such as mobile phone used in the case of bipolar illness [16]). As we navigate this intersection of technology and mental health, it becomes imperative to ensure that ethical considerations, patient safety, and data privacy remain paramount [17].

Building upon this foundation, our research introduces a novel application that further pushes the boundaries of what conversational agents can achieve in mental health care. By integrating cutting-edge ASR and NLP technologies, we have developed a system that not only understands and processes complex human speech but also responds in a contextually sensitive manner, thereby providing a more nuanced and effective therapeutic interaction. Recognizing the sensitive nature of mental health data, our approach is grounded in stringent ethical standards, ensuring the utmost respect for patient confidentiality and security. This commitment to ethical research practice is woven throughout our study, ensuring that the advancements we present are not only scientifically robust but also ethically sound, paving the way for a new era of responsible AI in mental health care.

2. Methods

In the pursuit of advancing the field of conversational AI within the context of online therapy applications, our study meticulously documents the methods and processes that were integral to our research. We begin by detailing the data preprocessing techniques and the datasets that laid the groundwork for our Automatic Speech Recognition (ASR) system. Following this, we describe the model selection criteria and training processes that were pivotal in developing a robust NLP framework. The subsequent sections delve into the user interaction dynamics, the technological frameworks employed, and the design principles that guided the creation of our conversational AI system. Finally, we outline the rigorous evaluation metrics that served to quantify the performance of our system.

2.1. Materials

The ASR system is integrated into a mobile application designed for online therapy, serving as a conversational agent to facilitate user interaction. It aids in various tasks, such as navigating the app, scheduling, modifying, postponing, or canceling appointments, and offers the option to dictate entries into a digital diary. The system's deployment in this context aims to enhance user experience

 doi:10.20944/preprints202402.1585.v1

by providing an intuitive and seamless interface, thereby reducing barriers to effective therapy engagement.

2.1.1. Data Preprocessing

We established an effective ASR system, as described in Figure 1, by meticulously preprocessing audio data, as recommended by [18]. We converted a variety of audio formats, such as mp3, mp4, and flac, to the widely compatible .wav format using the ffmpeg multimedia framework. We chose this format for its broad compatibility with numerous ASR models. Following this conversion, we standardized the audio data to a sample rate of 16000 Hz using the torch audio library [19], which is essential for consistent model training.

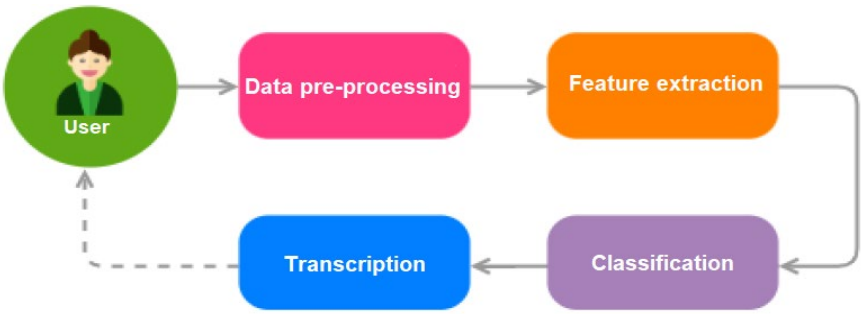

Figure 1. ASR process.

2.1.2. Dataset

The Mozilla Common Voice dataset was instrumental in training our ASR system. This open-source corpus is a rich collection of transcribed voice data that enables the development and benchmarking of ASR systems in a variety of languages [20]. For the purposes of our study, we focused on the French subset of the dataset, which provided a vast and diverse range of accents, dialects, and speaking styles, essential for developing a robust model for the French language. The dataset's structure, with its clear labelling and extensive metadata, facilitated a streamlined training process and allowed for efficient fine-tuning of the ASR model to the nuances of spoken French.

Acoustically, the Mozilla Common Voice dataset encompasses recordings from numerous environments, reflecting real-world scenarios where a user might interact with the therapy app. This variety in the audio data ensures that our ASR system is well-equipped to recognize and process speech in different acoustic settings, thereby improving its performance and reliability.

2.1.3. Archetypal Selection and Training

In our pursuit to address the challenges inherent to Automatic Speech Recognition (ASR), we turned to NVIDIA's NeMo project, a cutting-edge open-source platform designed specifically for ASR and other neural network tasks [21]. NeMo's repository boasts a plethora of pre-trained models, each tailored for specific applications and challenges within the realm of vocal AI.

After a rigorous evaluation of available models, our choice gravitated towards a specific pre-trained model, which has been previously recognized in the literature for its exemplary performance in speech-to-text conversion tasks [3]. This model, built on state-of-the-art architectures and training methodologies, promised a blend of accuracy and efficiency, making it an ideal candidate for our research objectives.

To further enhance the model's transcription capabilities, we incorporated the Connectionist Temporal Classification (CTC) algorithm (Figure 2) [22]. The CTC algorithm plays a pivotal role in aligning temporal sequences in audio data with their corresponding transcriptions, a challenge that is non-trivial given the variable speed and cadence of human speech. During the training phase, the CTC loss function was employed, serving as a guiding metric to iteratively refine and optimize the

 doi:10.20944/preprints202402.1585.v1

neural network. This ensured that the final model was adept at producing transcriptions that were not only accurate but also temporally coherent with the input audio.

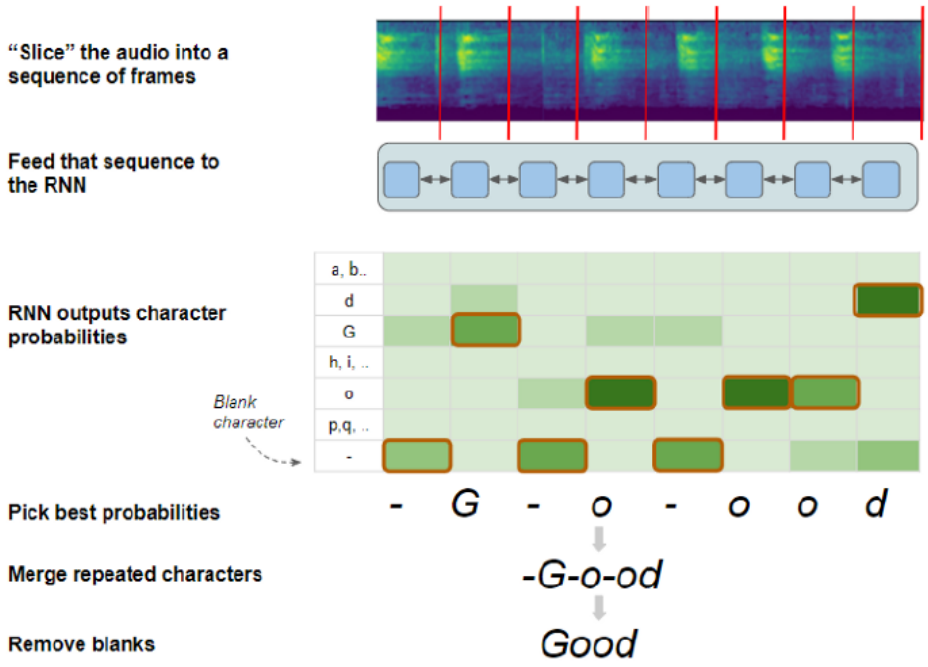

Figure 2. CTC decoding algorithm.

2.1.4. Language Modeling

Language modeling is a paramount in enhancing ASR systems, aiming to elevate transcription accuracy and minimize the word error rate (WER). Throughout the development of our project, we employed an N-gram language model (LM), which was trained on extensive textual data. This LM was then integrated with beam search decoding to ascertain the most probable transcription outcomes. The beam search decoders in NeMo are compatible with LMs trained using the KenLM library, allowing for a seamless fusion of acoustic and language modelling [23].

The beam search algorithm is a heuristic search strategy that expands multiple tokens at each position within a sequence. It can consider any number of 'N' best alternatives through an hyperparameter known as beam width. For instance, with a beam width set to 2, the algorithm selects the two most probable characters at each sequence position, branching out and combining probabilities to form the most likely sequences until an "<END>" token is encountered, thus determining the best transcription path.

The N-gram LM is particularly effective when used in conjunction with beam search decoders atop ASR models, as it refines the candidate outputs. The beam search decoder incorporates scores from the N-gram LM into its scoring calculations as follows:

$$final\_score = acoustic\_score + beam\_alpha \times lm\_score + beam\_beta \times seq\_length$$

Here, $acoustic\_score$ represents the prediction by the acoustic encoder, and lm_score is the estimate from the LM. The parameter $beam\_alpha$ dictates the weight given to the N-gram LM, influencing the balance between language and acoustic modeling. A higher $beam\_alpha$ indicates a stronger reliance on the LM, whereas $beam\_beta$ acts as a penalty term to account for sequence length in the scoring. Negative values for $beam\_beta$ penalize longer sequences, prompting the decoder to favour shorter predictions, while positive values bias towards longer candidate sequences.

This careful calibration of parameters, as depicted in Figure 3, is essential for fine-tuning the language model's performance, ensuring that the ASR system not only predicts with high precision but also reflects the inherent variability of human speech [23,24].

 doi:10.20944/preprints202402.1585.v1


```
return nemo_asr.modules.BeamSearchDecoderwWithlM(
    vocab=list(self.model.decoder.vocabulary),
    beam_width=16,
    alpha=2, beta=1.5,
    Im_path="./mls_Im_french/3-gram_lm.arpa",
    num_cpus=max(os.cpu_count(), 1),
    input_tensor=False)
```

Figure 3. Configuration of Beam Search Decoder with N-gram Language Model.

2.1.5. Model Architecture

We selected QuartzNet 15x5, a derivative of the Jasper architecture known for its robust performance in speech recognition tasks [21]. This specific variant of QuartzNet, composed of 79 layers with 5 blocks repeated 15 times and enriched by 4 additional convolutional layers, boasts 18.9 million parameters. Its convolutional design, trained using Connectionist Temporal Classification (CTC) loss, is particularly effective at capturing the intricacies of complex speech patterns due to its multiple blocks with residual connections. Recognizing the need for a model attuned to the nuances of the French language, the QuartzNet model was fine-tuned from English language to French using French portion of Common Voice from Mozilla (MCV)[25]. This dataset was selected for its wide range of accents and dialects, which provided the diversity necessary to train a more robust and versatile ASR system for the intricacies of spoken French.

2.2. Tasks and design

Within the burgeoning field of digital therapeutics, the advancement and refinement of Natural Language Processing (NLP) technologies are of critical importance. NLP systems serve as the foundational framework that facilitates sophisticated human-computer dialogue, a core component that is indispensable in the context of online therapy applications. The efficacy of these platforms is heavily reliant on the clarity and precision of communication, as these attributes are directly correlated with the user's experience and the therapeutic efficacy [11].

2.2.1. Conceptual Foundation

At the core of our research lies the intricate domain of Natural Language Processing (NLP), a subfield of AI that facilitates human-computer communication by enabling machines to comprehend and generate human language [26]. Within this domain, we delved into two crucial sub-disciplines.

- *Natural Language Understanding (NLU):* This facet of NLP focuses on converting user input into a structured format that algorithms can interpret, thereby discerning the underlying intent and entities in a given text [26].
- *Natural Language Generation (NLG):* Contrasting NLU, NLG is concerned with formulating coherent responses in natural language based on the machine's understanding [27].

As our research progresses, our primary objective is to refine the NLP components. The aim is to seamlessly integrate them, producing a holistic conversational AI system (Figure 4) that stands as a testament to the potential of NLP in revolutionizing voice-assisted systems [28–32].

 doi:10.20944/preprints202402.1585.v1

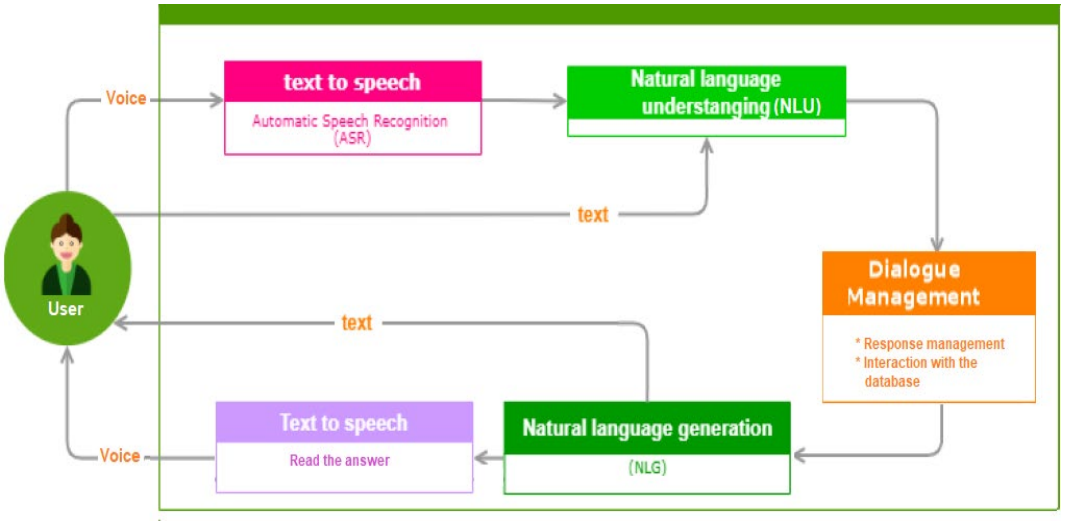

Figure 4. Voice Assistant flowchart.

2.2.2. User Interaction Dynamics

To ensure a seamless interaction between the user and the voice assistant, we delved into the principles of user intent and entities. For instance, a command like "take an appointment" translates to an intent, termed "prendreRDV" in our system. Concurrently, entities within these intents, such as time, were meticulously identified and labelled [33].

2.2.3. Rasa Architectural Components

In our research, we actively explored the Rasa architecture, focusing on the NLU Pipeline and Dialogue Management, to guarantee precise recognition of user intents and appropriate actions within conversations. Our in-depth examination of the Rasa framework's components included:

- *NLU Pipeline:* responsible for intent classification, entity extraction, and response generation [34]. It processes user inputs through a trained model, ensuring accurate intent recognition.
- *Dialogue Management:* discerns the optimal subsequent action in a conversation based on the immediate context [35].
- *Tracker Stores, Event Brokers, Model Storage, and Lock Stores:* collectively ensure the efficient storage of user interactions, integration with external services, and maintenance of message sequencing.

Building on the modular architecture of Rasa, we structured our project to encapsulate the full spectrum of conversational AI capabilities. The project is organized into several key files and directories, each with a specific role:

- domain.yml: a central configuration file that defines all the elements that the assistant can understand and produce. It includes:
  - Responses: The set of utterances the assistant can use in response to user inputs.
  - Intents: The classifications of user inputs that help the assistant interpret the user's intentions.
  - Slots: Variables that store information throughout the conversation, maintaining context and state.
  - Entities: Information extracted from user inputs that can be used to personalize interactions.
  - Forms & Actions: These enable the assistant to perform tasks and carry out dynamic conversations based on the dialogue flow.
- Config.yml: Specifies the machine learning model configurations, guiding the natural language understanding and dialogue management processes.
- data directory: Contains the training data that the assistant uses to learn and improve its understanding and dialogue management with nlu.yml for intent and entity examples, stories.yml for conversational paths, and rules.yml for dialogue policies.

 doi:10.20944/preprints202402.1585.v1

The Rasa framework's flexibility is exemplified by its ability to adapt to various conversational scenarios, making it an invaluable tool for our research and development efforts. By leveraging Rasa, we have crafted an assistant that is not only proficient in language understanding and generation, but also adept at managing complex conversational flows and maintaining context across interactions.

2.2.4. Data Preparation and Model Implementation

a) Conversational Design and Objective Identification

Central to our research was the principle of 'Conversation Design', which entailed structured planning of potential interactions, user profiling, understanding assistant objectives, and documenting typical user conversations [36].

b) Data Acquisition and Conversation Simulation

Lacking historical interaction logs, we simulated human-bot interactions, leveraging insights from domain experts and our customer service team [37]. We specifically employed the fr_core_news_sm model from spaCy, version 3.0. This choice was based on preliminary validation tests that demonstrated its superior performance in understanding and processing the French language compared to other available models.

c) NLU Pipeline and Language Model Choices

The fr_core_news_sm model, an efficient component of the spaCy library, was integral to our NLU pipeline. Its pre-trained embeddings were crucial for the linguistic analysis tailored to our project's needs, aligning with methodologies proven in health sector research [28–32]. The configuration of our NLP pipeline, optimized for our simulated dataset, is presented in Figure 5.

```yaml
pipeline:
  - name: SpacyNLP
    model: "fr_core_news_sm"
  - name: SpacyTokenizer
  - name: SpacyFeaturizer
    "pooling": "mean"
  - name: LexicalSyntacticFeaturizer
  - name: CountVectorsFeaturizer
  - name: CountVectorsFeaturizer
    analyzer: "char_wb"
    min_ngram: 2
    max_ngram: 4
  - name: DIETClassifier
    epochs: 150
  - name: DucklingEntityExtractor
    dimensions: ["time"]
```

Figure 5. NLU pipeline.

d) Text Tokenization and Featurization

We transformed our textual data into tokens suitable for machine interpretation by employing the SpacyTokenizer, which uses the linguistic annotations from the "fr_core_news_sm" model [28–32]. Following tokenization, we utilized the SpacyFeaturizer to generate dense word embeddings, where a mean pooling strategy was applied to create aggregated phrase representations. To encompass a wider range of linguistic attributes, we integrated two variations of the CountVectorsFeaturizer, capturing both word and character-level n-grams, thereby enhancing our model's ability to understand nuanced language patterns.

 doi:10.20944/preprints202402.1585.v1

e) Part-of-Speech Tagging and Intention Classification

Subsequent to feature extraction, we deployed the Dual Intent and Entity Transformer (DIET) classifier within the Rasa framework for intention classification and entity extraction. This classifier was chosen for its capability to perform both tasks concurrently, which is essential for understanding the intricacies of natural language in conversation. We further augmented our entity recognition capabilities with the inclusion of the DucklingEntityExtractor, enabling our model to reliably interpret various data formats and entities such as dates, times, and numerical values.

Intent Definitions and Training Data: A pivotal step in our NLU pipeline was the definition and categorization of intents. We meticulously compiled a dataset of utterances for each intent to facilitate robust training. Table 1 presents the intents recognized by our system, their definitions, associated entities, and the number of training examples for each.

Table 1. Intent Definitions and Training Data.

Intent	Definition	Entities	Training Data Count
goodbye	User wishes to say farewell	-	8
greet	Greetings	-	8
affirm	User confirms something	-	9
deny	User refuses or denies something	-	4
informApp	User seeks information about the application	-	14
informPacks	User inquires about the application's packages	-	17
bot_challenge	User asks if they are speaking to a bot or a human	-	3
prendreRdv	User requests an appointment	time	41
changerRdv	User requests to change their appointment	time	14
annulerRdv	User requests to cancel their appointment	-	18
raterRdv	User missed their appointment	-	6
informerRdv	User inquires about confirmed appointments	-	11
info_date	User asks for a date of the appointment	time	9
IdK	User responds with 'I don't know'	-	3
ageUser	User provides their age	age	9
raisonEmotion	User responds due to an undesirable emotion	-	4
entenduApp	User responds how they heard about the service	-	9
emotion_therapy	User explains why they need therapy	-	16
gerer_sentiment	User describes how they manage their feelings	-	6
out_of_scope	Intent for text that our assistant does not cover initially	-	6

f) Dialogue Management

We leveraged Rasa's core capabilities to decode and manage the flow of conversations. By utilizing a curated set of stories and rules as our training data, we empowered the assistant to accurately predict and execute the most appropriate action in response to user inputs during conversations.

g) Forms in Conversations

We integrated forms as a fundamental component of our conversational design to streamline specific tasks. These forms were crucial in efficiently handling user requests for scheduling or rescheduling appointments, ensuring a smooth and intuitive conversational experience.

2.2.5. Data Management and System Architecture

We architected the application to facilitate robust data handling and user interaction. Selecting Firebase as our NoSQL database platform, we capitalized on its scalability and real-time synchronization features (Figure 6). The database architecture comprises two primary collections:

 doi:10.20944/preprints202402.1585.v1

'conversation' for interaction logs and 'RDV' for appointment management, optimizing data retrieval and manipulation processes.

We constructed the ASR component utilizing NVIDIA's NeMo toolkit, which excels in capturing and transcribing speech with notable precision [19]. Simultaneously, the Rasa framework underpins our natural language understanding and dialogue management, interpreting user queries with a high degree of accuracy [42].

For the user interface, we employed the Flutter framework, renowned for its dynamic and responsive design capabilities, to craft an engaging user experience. The system's backend, orchestrated on a Flask server, efficiently handles requests and integrates with the frontend via HTTP protocols, ensuring swift and precise responses to user interactions.

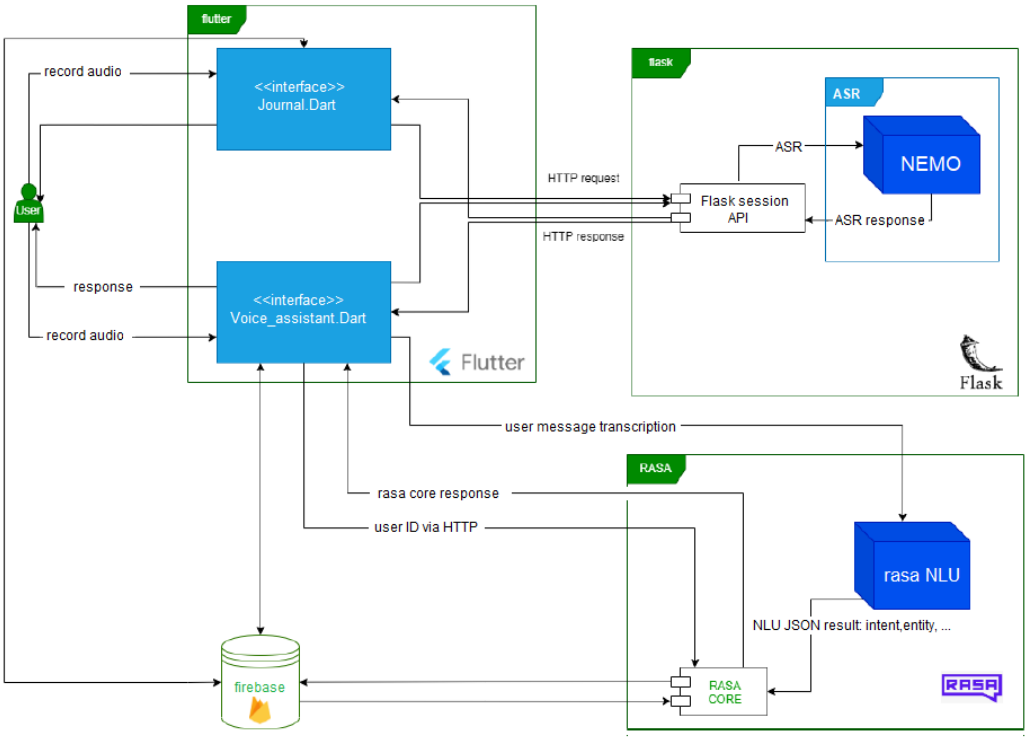

Figure 6. System Architecture - An overview of the system's infrastructure, illustrating the interplay between the ASR component, dialogue management, and the user interface.

2.3. Analysis

2.3.1. Evaluation of ASR performance

To rigorously assess the performance of our Automatic Speech Recognition (ASR) system, we employed the Word Error Rate (WER) as our primary evaluation metric. WER is a widely recognized and standard metric in the realm of speech recognition, providing a quantitative measure of the system's accuracy by comparing the system's predicted transcriptions against the actual or 'ground truth' transcriptions.

The WER is computed by determining the minimum number of operations – substitutions (S), insertions (I), and deletions (D) – required to align the predicted transcription with the ground truth. These operations are summed and then divided by the total number of words N in the ground truth transcription, as shown in the WER formula: $WER = (S + I + D)/N$. Here, N represents the number of words in the reference transcription, which serves as the normalization factor for the error rate. A lower WER indicates higher transcription accuracy, while a higher WER suggests potential areas of improvement in the ASR system.

The choice of WER as our evaluation metric is grounded in its ability to offer a comprehensive view of the system's performance. By accounting for all types of transcription errors, WER provides

 doi:10.20944/preprints202402.1585.v1

a holistic assessment, ensuring that we capture the full spectrum of discrepancies between the predicted and actual transcriptions. This approach is in line with the findings of [43], who highlighted the significance of WER in evaluating ASR systems, emphasizing its utility in pinpointing areas of potential refinement and optimization.

2.3.2. NLP System Evaluation

For the evaluation of intent recognition and entity extraction, we used a combination of "Train-Test Split" and "cross-validation" methods to assess the robustness of our models. Cross-validation was particularly emphasized due to its advantage in utilizing the dataset more effectively, providing a comprehensive view of the model's performance across different subsets of data.

For cross-validation, we partitioned the data into five folds, ensuring that each fold was a good representative of the whole. We then trained our model on four folds and validated it on the fifth, repeating this process five times so that each fold served as the validation set once. The performance metrics from each fold were then aggregated to give an overall performance measure.

The metrics for NLP system evaluation included precision, recall, and F1-score for both intent classification and entity recognition. These metrics were calculated using the scikit-learn library, which provided functions for model training, cross-validation, and performance evaluation.

2.3.3. Deployment and Database Integration

Following deployment, we meticulously monitored the system's performance metrics, including response times and user satisfaction rates. The integration of the ASR and NLP components into the Flask server facilitated a seamless data flow between the mobile application and backend services, yielding prompt and accurate responses to user queries. The deployment pipeline's effectiveness was evidenced by the automated testing and updating of models, ensuring continuous system optimization. We leveraged the Flask framework's capabilities to manage requests and maintain real-time communication with the frontend application [44].

User interactions, system responses, and performance metrics were logged systematically, providing a rich dataset for ongoing analysis and system refinement. This data-driven approach allowed us to iteratively enhance the system's accuracy and user experience, as reflected in the positive feedback from the application's user base.

2.3.4. Ethical Considerations and Data Privacy

In this study, a minimal data collection approach was adopted to prioritize user privacy and align with ethical standards. Only user email addresses and pseudonyms were collected, with no additional personal identifiers. This measure ensures a high degree of anonymity, reducing the risk of personal data exposure. Before participating, users were informed about the study's objectives, the extent of data usage, and their right to withdraw at any point, from which informed consent was obtained. All data were handled in accordance with the European General Data Protection Regulation (GDPR), emphasizing data minimization, integrity, and confidentiality. The use of emails and pseudonyms was strictly for communication and personalization within the study, ensuring that users' privacy was maintained. Despite the limited nature of personal data collection, our commitment to ethical standards and user privacy remained paramount throughout the research process.

3. Results

3.1. ASR System Performance

The accuracy of the model was evaluated using the development set from French Mozilla Common Voice (MCV) dataset. The Word Error Rate (WER) served as a primary metric, reflecting the percentage of errors within the model's transcriptions. The model obtains a WER of 14% on the development set of MCV dataset. Using this model, our system demonstrates a commendable level

 doi:10.20944/preprints202402.1585.v1

of accuracy in transcribing spoken French. This result is particularly significant given the diversity of accents, dialects, and speaking styles present in the Common Voice data, which closely mimics the variability encountered in real-world scenarios. The proficiency exhibited by our system suggests its suitability for practical applications, such as facilitating user interaction with a conversational agent in an online therapy mobile application.

3.2. NLP System Evaluation

Subsequent to ASR transcription, the NLP system was subjected to a thorough evaluation, focusing on its precision, recall, and F1-score for various intents and entities. These metrics are pivotal as they directly influence the user experience by determining the system's ability to comprehend and respond to user inputs with precision.

Figure 7 presents the confusion matrix for intent recognition, providing insight into system's ability to classify user intents correctly.

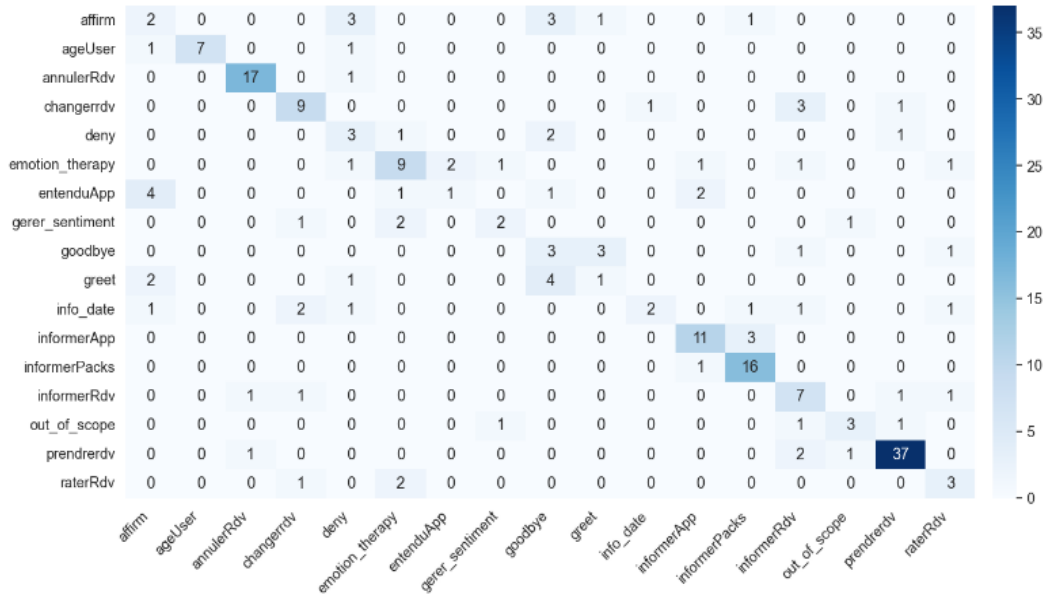

Figure 7. Intent Recognition Confusion Matrix.

Figure 8 depicts the confusion matrix for the DIET classifier, highlighting the system's performance in entity extraction, and provides a granular view of the system's performance, showcasing areas of strength and those necessitating further optimization. The Table 2 presents a comprehensive breakdown of precision, recall, and F1-scores, supplemented by the 'Support' column which indicates the volume of samples for each category within the test dataset. The 'Confused With' column offers valuable insights into the most frequent misclassifications, guiding potential enhancements to the model.

 doi:10.20944/preprints202402.1585.v1

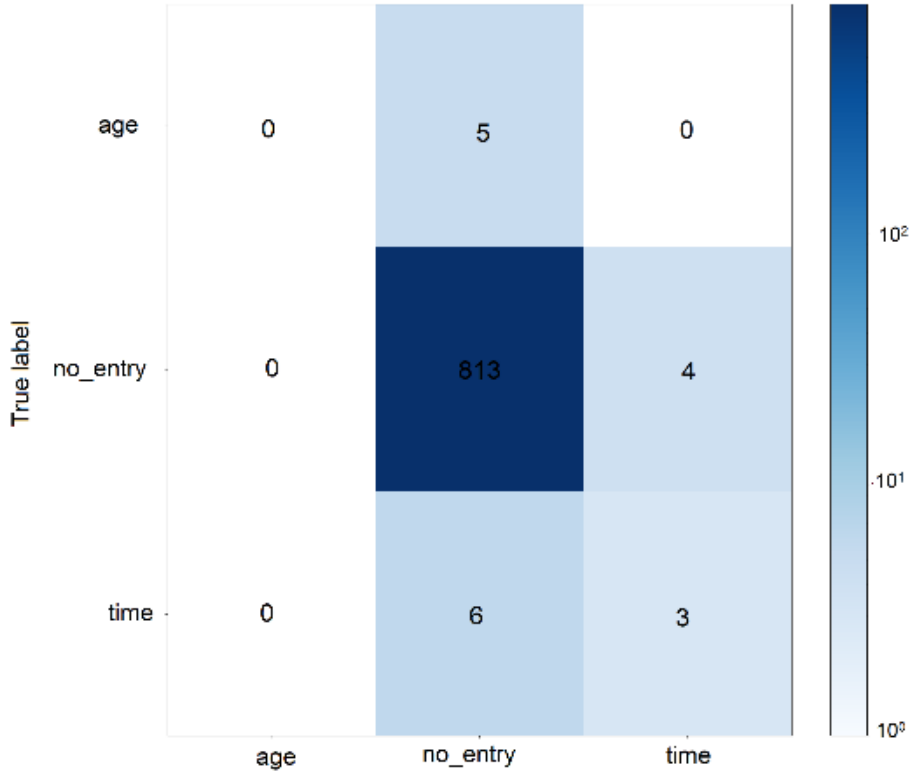

Figure 8. DIET Classifier Confusion Matrix.

For instance, the intent 'Prendre Rdv' demonstrated exceptional precision and recall, both surpassing the 90% threshold, denoting its reliable recognition within the system. Conversely, the intent 'greet' manifested lower metrics, signifying a domain where the classifier's discernment could be improved.

Table 2. Detailed Intent and Entity Performance Metrics.

Intent/ Entity	Precision	Recall	F1-Score	Support	Confused With
time	42.86%	33.33%	37.50%	9	-
age	0.00%	0.00%	0.00%	5	-
Gerer sentiment	50.00%	33.33%	40.00%	6	emotion_therapy (2), out_of_scope (1)
Informer Rdv	43.75%	63.64%	51.85%	11	raterRdv (1), annulerRdv (1)
Prendre Rdv	90.24%	90.24%	90.24%	41	informerRdv (2), out_of_scope (1)
Emotion therapy	60.00%	56.25%	58.06%	16	entenduApp (2), raterRdv (1)
goodbye	23.08%	37.50%	28.57%	8	greet (3), informerRdv (1)
raterRdv	42.86%	50.00%	46.15%	6	emotion_therapy (2), changerRdv (1)
greet	20.00%	12.50%	15.38%	8	goodbye (4), affirm (2)
deny	27.27%	42.86%	33.33%	7	Goodbye (2), emotion_therapy (1)
info_date	66.67%	22.22%	33.33%	9	changerRdv (2), raterRdv (1)
Informer Packs	76.19%	94.12%	84.21%	17	informerApp (1)
affirm	20.00%	20.00%	20.00%	10	goodbye (3), deny (3)
informer App	73.33%	78.57%	75.86%	14	informerPacks (3)
out of scope	60.00%	50.00%	54.55%	6	informerRdv (1), prendreRdv (1)
annuler Rdv	89.47%	94.44%	91.89%	18	deny (1)
entendu App	33.33%	11.11%	16.67%	9	affirm (4), informerApp(2), emotion therapy (1)

 doi:10.20944/preprints202402.1585.v1

changer Rdv	64.29%	64.29%	64.29%	14	informerRdv (3), info_date (1)
ageUser	100.00%	77.78%	87.50%	9	affirm (1), deny (1)
Overall	64.24%	63.64%	62.75%	209	-

3.3. Error Analysis and Model Confidence

The error analysis is instrumental in delineating the model's limitations and informing subsequent iterations. Figure 9 Intent Prediction Confidence Distribution, portrays the variance in prediction confidences across different intents, thus reflecting the model's certainty in its classifications.

- *Model Performance:* The model assigns confidence scores that range broadly, from high confidence, such as a score of 0.995 for correctly predicting "salut" as "greet," to moderate confidence levels, such as a score of 0.827 for classifying "je veux que la date soit demain" as "changerRdv." This range indicates the model's varied levels of certainty in its predictions.
- *Accuracy and Misclassification:* Notably, the model sometimes misclassifies intents with substantial confidence. For example, "je refuse" is misclassified as "emotion_therapy" with a confidence score of 0.328, highlighting a clear area for model improvement.
- *Confidence Score Distribution:* The distribution of confidence scores is indicative of the model's predictive certainty. A concentration of high scores would imply a high degree of certainty in its predictions, whereas a more dispersed set of scores could signal the necessity for further model calibration.

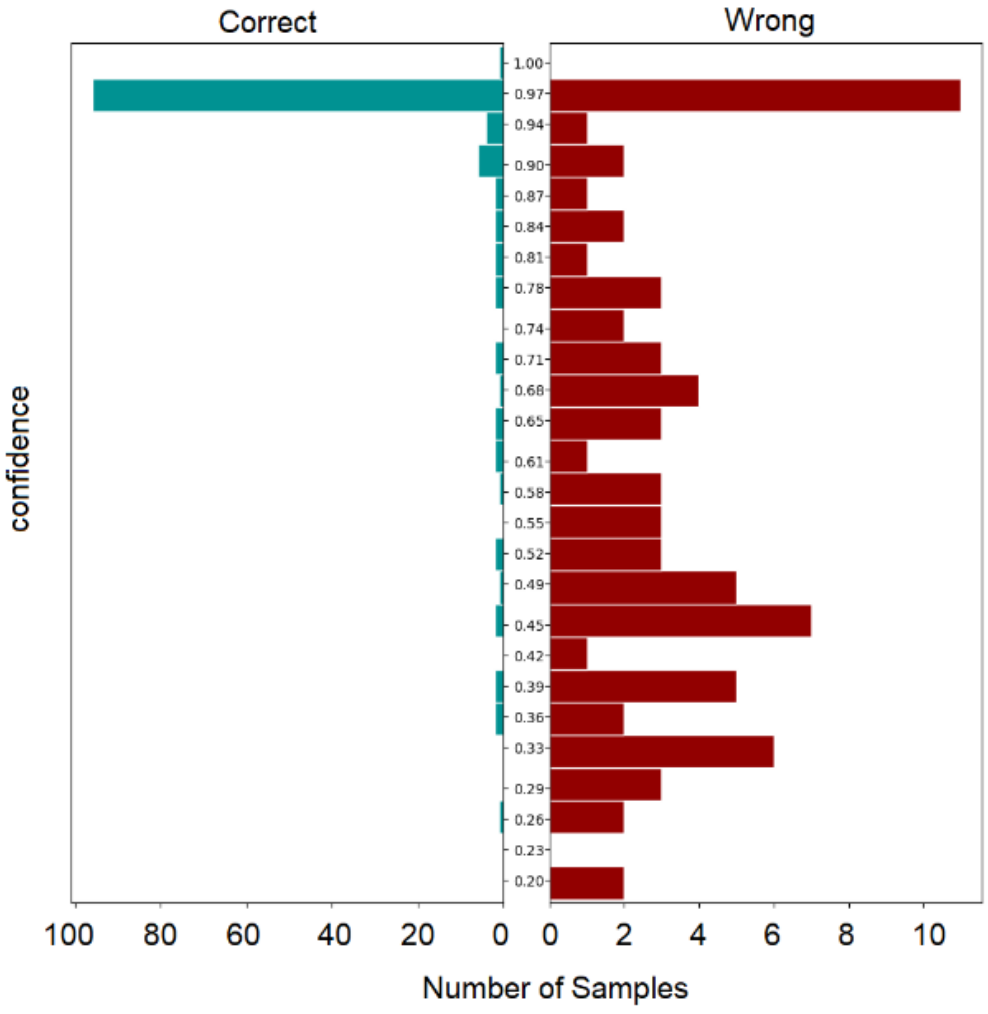

Figure 9. Intent Prediction Confidence Distribution.

3.4. System Integration and Deployment

 doi:10.20944/preprints202402.1585.v1

The integration of the ASR and NLP systems into a cohesive voice-assistant application is substantiated by its deployment within an operational setting. User feedback and interaction logs have been overwhelmingly positive, with users particularly highlighting the system's intuitive conversational interface. This interface has significantly enhanced user engagement and experience, facilitating easier navigation and interaction within the online therapy platform.

- *User Interaction:* The application's conversational agent has successfully assisted users in managing appointments, navigating the platform, and utilizing the diary feature through voice commands. This has been particularly beneficial for users with varying levels of technical proficiency.
- *Continuous Improvement:* Regular user feedback and system interaction logs are critical for ongoing refinement, ensuring the system remains responsive to user needs and operational demands.

In reflecting on the system's deployment and user feedback, we conducted a further analysis of the system's performance metrics to identify opportunities for enhancement. The ASR system, while performing at a commendable level, displayed a tendency for phonetic misinterpretations, particularly in challenging acoustic environments. The NLP system's precision in intent recognition, while generally high, showed areas for improvement in handling complex language structures and expressions. These insights were paralleled by user feedback, which, despite being overwhelmingly positive, suggested areas where the conversational interface could become more intuitive. This feedback loop from real-world application is invaluable, providing direct guidance for the continuous refinement of our system. It assures us that while our system is on the right path, there is a journey ahead toward achieving seamless human-computer linguistic interaction within the therapeutic context.

4. Discussion

The integration of Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) technologies into mental health services is an emerging area that presents distinct challenges and opportunities. Our study contributes to this domain by implementing a French-language ASR system within an online therapy application, addressing a gap in current research. The Word Error Rate (WER) of 14% on the Mozilla Common Voice development set, while not leading the industry, represents a significant step forward given the limited development of French ASR datasets and models, particularly for the nuanced context of therapeutic dialogue.

In comparison with the broader healthcare sector, where ASR systems are increasingly utilized, such as in clinical documentation and patient care, our system's performance is modest. This is partly due to the predominance of English-centric research, which benefits from more advanced models and datasets. The scarcity of comprehensive French-language datasets, especially in specialized fields like mental health, is a notable limitation. The ' Mozilla Common Voice ' dataset, while valuable, underscores the need for more extensive resources that capture the full spectrum of French linguistic variations.

The potential benefits of ASR and NLP technologies in mental health care are significant, as they can provide support and enhance the accessibility of services. However, the sensitive nature of mental health dialogue requires a level of linguistic and emotional intelligence that our current system is still striving to achieve. This is particularly challenging in French, where resources to train such systems are less available.

The technical robustness of our ASR and NLP models was a primary focus, yet we recognize their current limitations. The system's WER of 14% is a testament to our progress with French ASR, but it also highlights the need for improvement. Most errors were phonetic misinterpretations, common in therapeutic dialogues due to their emotional and nuanced nature. We have begun addressing these through iterative refinements in model training and data preprocessing to better capture linguistic subtleties. Future versions will incorporate user feedback to reduce biases and improve the system's emotional intelligence.

 doi:10.20944/preprints202402.1585.v1

To move forward, it is imperative that future research focuses on the development of comprehensive French-language datasets that reflect the diversity of the mental health service users population. Collaborative efforts between technologists, clinicians, and linguists are crucial to create models that can accurately interpret the subtleties of therapeutic communication. Moreover, the development of adaptive models that can learn from user interactions and tailor responses to the specific linguistic and emotional context of therapy sessions will be key for future progress.

In conclusion, our work represents progress in French-language ASR for mental health applications, but it also clearly delineates the need for investment in the creation of rich datasets and the development of sophisticated models. Such advancements will not only enhance the accuracy of ASR systems, but will also ensure that they meet the critical needs of mental health support in the digital age.

As we acknowledge the progress made, it is crucial to highlight the ethical considerations inherent in applying ASR and NLP to mental health. Our study has maintained a steadfast commitment to the ethical deployment of these technologies, ensuring they serve as a complement rather than a substitute for human empathy and understanding. We recognize the potential for biases in our system, particularly given the unique linguistic intricacies of French as used in therapeutic settings. Efforts to mitigate these biases have been integral to our development process, and we continue to refine our algorithms to better capture the emotional valence and therapeutic intent of conversations. Additionally, we have implemented feedback mechanisms that allow continuous improvement of the system based on user interactions. These steps are essential in moving towards a more sensitive, accurate, and user-centered application of AI in mental health, which respects the nuances of human language and emotion. The future of our research will emphasize not only the technical refinement of ASR and NLP models, but also the cultivation of their ethical application, ensuring they align with the overarching goal of augmenting mental health support in a responsible and patient-centric manner.

5. Conclusions

In summary, our research has made a notable contribution to the field of ASR and NLP within the context of mental health services by developing a French-language ASR system tailored for an online therapy platform. Despite the challenges posed by the scarcity of robust French-language datasets, our system achieved a WER of 14% on the Mozilla Common Voice development set, demonstrating its potential utility in real-world therapeutic settings.

The significance of our work extends beyond the technical achievement of the ASR system's performance. It lays the groundwork for future innovations in the delivery of mental health services, where the nuances of language and the need for empathetic communication are paramount. By bridging the gap in French-language resources, we pave the way for more inclusive and accessible mental health care.

Looking ahead, the path is clear for the continued evolution of ASR and NLP technologies in healthcare. The development of comprehensive datasets and sophisticated models that can understand and respond to the complexities of human language and emotion is crucial. Our study serves as a stepping stone towards the realization of more effective and empathetic digital mental health services, and we anticipate that subsequent research will build upon our findings to further enhance the capabilities of ASR systems in this vital sector.

As we conclude, we reflect on the importance of interdisciplinary collaboration in advancing these technologies. The intersection of computational linguistics, clinical expertise, and user-centered design is the point where significant progress will be made. It is our hope that this research not only informs, but also inspires continued efforts to develop tools that support mental well-being in the digital age.

Author Contributions: Conceptualization, Mariem Jelassi, Houssem Ben Khalfallah and Jacques Demongeot; Methodology, Mariem Jelassi, Khouloud Matteli, Houssem Ben Khalfallah and Jacques Demongeot; Software, Khouloud Matteli and Houssem Ben Khalfallah; Validation, Mariem Jelassi and Khouloud

 doi:10.20944/preprints202402.1585.v1

Matteli; Formal analysis, Houssem Ben Khalfallah; Investigation, Mariem Jelassi, Khouloud Matteli and Houssem Ben Khalfallah; Resources, Mariem Jelassi; Data curation, Khouloud Matteli; Writing – original draft, Mariem Jelassi; Writing – review & editing, Mariem Jelassi, Houssem Ben Khalfallah and Jacques Demongeot.

References

1. F. Jelinek, "Statistical Methods for Speech Recognition," MIT Press. Accessed: Nov. 09, 2023. [Online]. Available: https://mitpress.mit.edu/9780262546607/statistical-methods-for-speech-recognition/
2. G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.
3. D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, PMLR, 2016, pp. 173–182. Accessed: Nov. 01, 2023. [Online]. Available: http://proceedings.mlr.press/v48/amodei16.html?ref=https://codemonkey.link
4. B. A. Shawar and E. Atwell, "Chatbots: are they really useful?," *Journal for Language Technology and Computational Linguistics*, vol. 22, no. 1, pp. 29–49, 2007.
5. A. C. Smith *et al.*, "Telehealth for global emergencies: Implications for coronavirus disease 2019 (COVID-19)," *J Telemed Telecare*, vol. 26, no. 5, pp. 309–313, Jun. 2020, doi: 10.1177/1357633X20916567.
6. T. Greenhalgh, J. Wherton, S. Shaw, and C. Morrison, "Video consultations for covid-19," *Bmj*, vol. 368. British Medical Journal Publishing Group, 2020. Accessed: Nov. 01, 2023. [Online]. Available: https://www.bmj.com/content/368/bmj.m998
7. D. M. Mann, J. Chen, R. Chunara, P. A. Testa, and O. Nov, "COVID-19 transforms health care through telemedicine: evidence from the field," *Journal of the American Medical Informatics Association*, vol. 27, no. 7, pp. 1132–1135, 2020.
8. A. Maier *et al.*, "PEAKS–A system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, no. 5, pp. 425–437, 2009.
9. T. W. Bickmore, L. M. Pfeifer, and M. K. Paasche-Orlow, "Using computer agents to explain medical documents to patients with low health literacy," *Patient education and counseling*, vol. 75, no. 3, pp. 315–320, 2009.
10. M. P. Turakhia *et al.*, "Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: The Apple Heart Study," *American heart journal*, vol. 207, pp. 66–75, 2019.
11. [11] Woebot. Accesseed: Jan. 10, 2024. [Online]. https://woebothealth.com/
12. A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, and J. B. Torous, "Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape," *Can J Psychiatry*, vol. 64, no. 7, pp. 456–464, Jul. 2019, doi: 10.1177/0706743719828977.
13. B. Inkster, S. Sarda, and V. Subramanian, "An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study," *JMIR mHealth and uHealth*, vol. 6, no. 11, p. e12106, 2018.
14. S. Schachter and J. Singer, "Cognitive, social, and physiological determinants of emotional state.," *Psychological review*, vol. 69, no. 5, p. 379, 1962.
15. G. M. Lucas, J. Gratch, A. King, and L.-P. Morency, "It's only a computer: Virtual humans increase willingness to disclose," *Computers in Human Behavior*, vol. 37, pp. 94–100, 2014.
16. T. W. Bickmore and R. W. Picard, "Establishing and maintaining long-term human-computer relationships," *ACM Trans. Comput.-Hum. Interact.*, vol. 12, no. 2, pp. 293–327, Jun. 2005, doi: 10.1145/1067860.1067867.
17. [17] T. Aubourg, J. Demongeot, F. Renard, H. Provost, and N. Vuillerme, "Association between social asymmetry and depression in older adults. A phone Call Detail Records analysis," *Scientific Reports*, vol. 9, pp. 13524, 2019, doi: 10.1038/s41598-019-49723-8.
18. S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.
19. A. van den Oord *et al.*, "WaveNet: A Generative Model for Raw Audio." arXiv, Sep. 19, 2016. Accessed: Nov. 01, 2023. [Online]. Available: http://arxiv.org/abs/1609.03499
20. V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210. Accessed: Nov. 01, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/ 7178964/
21. O. Kuchaiev *et al.*, "NeMo: a toolkit for building AI applications using Neural Modules." arXiv, Sep. 13, 2019. Accessed: Nov. 01, 2023. [Online]. Available: http://arxiv.org/abs/1909.09577
22. A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international*

conference on Machine learning - ICML '06, Pittsburgh, Pennsylvania: ACM Press, 2006, pp. 369–376. doi: 10.1145/1143844.1143891.
23. K. Heafield, "KenLM: Faster and smaller language model queries," in *Proceedings of the sixth workshop on statistical machine translation*, 2011, pp. 187–197. Accessed: Nov. 01, 2023. [Online]. Available: https://aclanthology.org/W11-2123.pdf
24. S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–394, 1999.
25. R. Ardila *et al.*, "Common Voice: A Massively-Multilingual Speech Corpus." arXiv, Mar. 05, 2020. Accessed: Nov. 01, 2023. [Online]. Available: http://arxiv.org/abs/1912.06670
26. J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, Jul. 2015, doi: 10.1126/science.aaa8685.
27. E. Reiter and R. Dale, "Building applied natural language generation systems," *Natural Language Engineering*, vol. 3, no. 1, pp. 57–87, 1997.
28. D. B. Dhiman, "Artificial Intelligence and Voice Assistant in Media Studies: A Critical Review," *Available at SSRN 4250795*, 2022, Accessed: Nov. 02, 2023. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4250795
29. R. S. Dinesh, R. Surendran, D. Kathirvelan, and V. Logesh, "Artificial Intelligence based Vision and Voice Assistant," in *2022 International Conference on Electronics and Renewable Systems (ICEARS)*, IEEE, 2022, pp. 1478–1483. Accessed: Nov. 02, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9751819/
30. J. N. Gupta, G. A. Forgionne, and M. Mora, "Intelligent decision-making support systems: foundations, applications and challenges," 2007.
31. B. Kadali, N. Prasad, P. Kudav, and M. Deshpande, "Home automation using chatbot and voice assistant," in *ITM Web of Conferences*, EDP Sciences, 2020, p. 01002. Accessed: Nov. 02, 2023. [Online]. Available: https://www.itm-conferences.org/articles/itmconf/abs/2020/02/itmconf_icacc2020_01002/itmconf_icacc2020_01002.html
32. D. Patel *et al.*, "An implementation framework and a feasibility evaluation of a clinical decision support system for diabetes management in secondary mental healthcare using CogStack," *BMC Med Inform Decis Mak*, vol. 22, no. 1, p. 100, Dec. 2022, doi: 10.1186/s12911-022-01842-5.
33. H. Chen, X. Liu, D. Yin, and J. Tang, "A Survey on Dialogue Systems: Recent Advances and New Frontiers," *SIGKDD Explor. Newsl.*, vol. 19, no. 2, pp. 25–35, Nov. 2017, doi: 10.1145/3166054.3166058.
34. A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017, Accessed: Nov. 02, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/7181-attention-is-all
35. I. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proceedings of the AAAI conference on artificial intelligence*, 2016. Accessed: Nov. 02, 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/9883
36. J. Cassell *et al.*, "Embodiment in conversational interfaces: Rea," in *Proceedings of the SIGCHI conference on Human factors in computing systems the CHI is the limit - CHI '99*, Pittsburgh, Pennsylvania, United States: ACM Press, 1999, pp. 520–527. doi: 10.1145/302979.303150.
37. M. F. McTear, "Spoken dialogue technology: enabling the conversational user interface," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 90–169, Mar. 2002, doi: 10.1145/505282.505285.
38. J. Delorme *et al.*, "Natural Language Processing for Patient Selection in Phase I or II Oncology Clinical Trials," *JCO Clinical Cancer Informatics*, no. 5, pp. 709–718, Dec. 2021, doi: 10.1200/CCI.21.00003.
39. M. Vincent, M. Douillet, I. Lerner, A. Neuraz, A. Burgun, and N. Garcelon, "Using deep learning to improve phenotyping from clinical reports," *Stud Health Technol Inform*, vol. 290, pp. 282–6, 2022.
40. M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," *To appear*, vol. 7, no. 1, pp. 411–420, 2017.
41. S. Bird, E. Klein, and E. Loper, "Natural Language Processing with Python [Book]." Accessed: Nov. 09, 2023. [Online]. Available: https://www.oreilly.com/library/view/natural-language-processing/9780596803346/
42. T. Bocklisch, J. Faulkner, N. Pawlowski, and A. Nichol, "Rasa: Open Source Language Understanding and Dialogue Management." arXiv, Dec. 15, 2017. Accessed: Nov. 02, 2023. [Online]. Available: http://arxiv.org/abs/1712.05181
43. A. C. Morris, V. Maier, and P. Green, "From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition," in *Eighth International Conference on Spoken Language Processing*, 2004.
44. M. Grinberg, Flask web development: developing web applications with python. O'Reilly Media, Inc., 2018.

 doi:10.20944/preprints202402.1585.v1

disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.