information

Article

Enhancing Personalized 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, Ecole Nationale des Sciences de l'Informatique, Manouba University, La Manouba 2010, Tunisia; mariem.jelassi@ensi-uma.tn (M.J.); houssem.ben-khalfallah@univ-grenoble-alpes.fr (H.B.K.)
[2] ESEN, Manouba University, La Manouba 2010, Tunisia; khouloud.matteli@esen.tn
[3] AGEIS Laboratory, University Grenoble Alpes, 38700 La Tronche, France
[*] Correspondence: jacques.demongeot@univ-grenoble-alpes.fr; Tel.: +33-615960440

Abstract: Automatic speech recognition (ASR) and natural language processing (NLP) play key roles in advancing human–technology interactions, particularly in healthcare communications. This study aims to enhance French-language online mental health platforms through the adaptation of the QuartzNet 15 × 5 ASR model, selected for its robust performance across a variety of French accents as demonstrated on the Mozilla Common Voice dataset. The adaptation process involved tailoring the ASR model to accommodate various French dialects and idiomatic expressions, and integrating it with an NLP system to refine user interactions. The adapted QuartzNet 15 × 5 model achieved a baseline word error rate (WER) of 14%, and the accompanying NLP system displayed weighted averages of 64.24% in precision, 63.64% in recall, and an F1-score of 62.75%. Notably, critical functionalities such as 'Prendre Rdv' (schedule appointment) achieved precision, recall, and F1-scores above 90%. These improvements substantially enhance the functionality and management of user interactions on French-language digital therapy platforms, indicating that continuous adaptation and enhancement of these technologies are beneficial for improving digital mental health interventions, with a focus on linguistic accuracy and user satisfaction.

Keywords: conversational AI; automatic speech recognition (ASR); natural language processing (NLP); online therapy platforms; AI in mental healthcare; French-language ASR; French NLP

check for updates

Citation: Jelassi, M.; Matteli, K.; Ben Khalfallah, H.; Demongeot, J. Enhancing Personalized Mental Health Support Through Artificial Intelligence: Advances in Speech and Text Analysis Within Online Therapy Platforms. *Information* **2024**, *15*, 813. https://doi.org/10.3390/info15120813

Academic Editor: Katsuhide Fujita

Received: 22 October 2024
Revised: 6 December 2024
Accepted: 7 December 2024
Published: 18 December 2024

1. Introduction

In the realm of digital innovation, automatic speech recognition (ASR) and natural language processing (NLP) have emerged as transformative forces in redefining human–technology interactions, especially within the healthcare sector, where seamless communication is essential. From the early systems constrained by limited vocabulary recognition to advanced models capable of deciphering a wide range of languages and dialects, the field has experienced significant evolution. This progress is thoroughly documented in seminal works like that of Dahl et al. (2011), who investigated context-dependent pre-trained deep neural networks for speech recognition, highlighting the remarkable advancements in ASR technologies [1]. These advancements transcend mere technical achievements, notably enhancing patient–provider interactions and overall psychological well-being [2,3].

The emergence of conversational agents, powered by sophisticated ASR and NLP technologies, marks a significant paradigm shift from simple chatbots to complex systems capable of engaging in human-like dialogues. This shift is reflected in the expanding applications of these agents from customer service to offering detailed medical advice [4] and deployment in mental health care [5]. The integration of these intelligent systems into healthcare practices not only promises enhanced operational efficiency, but also improved accessibility, particularly in mental health care, where they deliver interventions grounded in cognitive behavioral therapy principles to effectively manage symptoms of conditions such as depression and anxiety [6,7].

Voice-assisted systems have evolved significantly due to advances in NLP and ASR technologies. Modern ASR systems, unlike their predecessors, now handle diverse dialects and accents with much greater accuracy, enabling more inclusive and personalized interactions. Additionally, they maintain conversational context and can adapt to the emotional tone of the user, features especially relevant in mental health contexts where empathy and context are paramount. These advancements have transformed voice-assisted systems from simple command-based interfaces into sophisticated conversational agents capable of supporting therapeutic engagement and personalized interventions.

A key development in healthcare NLP is the transition towards end-to-end models, moving away from traditional segmented processing approaches. This evolution is supported by the work of Cho et al. (2014), who showcased the efficacy of RNN encoder–decoder models for phrase representation learning, contributing significantly to the foundational understanding of end-to-end models' potential [8]. Unlike conventional NLP systems that rely on distinct processing stages, end-to-end models employ deep learning to process and respond to inputs in a unified manner, enhancing interaction experiences by offering more coherent and contextually relevant responses [9,10].

Despite these significant advancements and the potential of ASR and NLP in healthcare, numerous challenges remain. The variability in speech patterns, dialects, and accents poses considerable challenges for ASR systems, potentially affecting their accuracy and effectiveness across diverse patient populations. Moreover, NLP systems, while proficient at processing structured language, often struggle with the subtleties of human communication, such as sarcasm, humor, and indirect speech, frequently encountered in therapeutic settings. The integration of these technologies into clinical practice faces hurdles, including issues surrounding data privacy, security, and the need for robust consent mechanisms, given the sensitive nature of mental health information. The risk of algorithmic biases, resulting from training data that may not fully represent diverse patient demographics, raises ethical concerns and the potential for disparities in care. Additionally, user acceptance and trust are essential; both patients and healthcare providers must feel comfortable with and trust these technologies for their successful deployment, overcoming resistance to change, digital literacy gaps, and concerns about the depersonalization of care.

In the context of French-speaking regions, the linguistic diversity and cultural nuances of the French language present unique challenges for the successful application of ASR and NLP in healthcare. The complexity of French dialects, from the European French language to various African, Canadian, and Ultramarine (notably in the Pacific and Caribbean) necessitates a nuanced approach to language modeling to ensure systems are responsive to the broad spectrum of linguistic variations. The challenge is compounded by the need for comprehensive and representative datasets that encapsulate this diversity, essential for training robust models. The work by Mancone et al. (2023) illustrates the potential of voice assistants like Alexa in eliciting empathy and engagement in psychological assessments while maintaining sound psychometric properties, indicating a promising direction for ASR and NLP applications in mental health, despite the challenges [11]. Ethical considerations, particularly in mental health applications, where sensitivity and privacy are paramount, are of equal importance. Ensuring data privacy involves strict adherence to regulatory frameworks like the General Data Protection Regulation (GDPR) in the European Union, which governs the use and protection of personal data. Additionally, the potential for algorithmic biases poses a significant ethical challenge, necessitating the development of models that are linguistically inclusive and free from biases that could lead to disparities in care.

This research endeavors to unveil the transformative potential of ASR and NLP in creating an interactive, personalized, and accessible online therapy platform, specifically catering to the nuances of the French language and culture. Through a comprehensive evaluation framework, this study aims to assess the platform's usability, therapeutic impact, and overall user satisfaction, addressing the unique challenges presented by the French linguistic and cultural landscape. Collaborative pilot studies with healthcare institutions, in line with the high-performance medicine models proposed by Topol (2019), will further validate the clinical relevance and efficacy of this AI-enhanced platform, marking a significant advancement in digital mental health solutions [12].

More generally, ASR is now capable of transcribing multiple languages, dialects, and accents with a remarkable precision [13]. Such advancements are not mere technological marvels; they hold profound implications for sectors like healthcare, where communication is fundamental [14,15]. Conversational agents, underpinned by ASR and enriched with 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 [15]. 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 epicenter. In oncology, these technologies have transitioned from experimental tools to essential components, offering patients comprehensive guidance on treatments, potential side effects, and post-treatment recovery [16]. For the elderly, whose increasing population is often challenged by rapid technological advancements, ASR-integrated devices have become indispensable, monitoring daily activities and ensuring timely medical interventions [17]. 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 monitoring of vital signs [18]. Their transformative impact extends to rehabilitation, aiding patients with speech and mobility impairments [19], and to pediatric care, where they simplify medical terminology for young patients [20]. 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 [21].

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.

Having outlined the technological advancements and the broader implications of ASR and NLP in healthcare, we now delve deeper into how these innovations are specifically reshaping mental health practices, addressing the ethical considerations and the need for culturally sensitive implementations that are essential for their success.

2. Background

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 [22]), employ principles from cognitive behavioral therapy (CBT) to engage users, showing significant efficacy in mitigating symptoms of depression and anxiety [23]. Numerous studies underscore the potential of these agents in delivering psychological interventions, with some rivaling the effectiveness of human therapists [24]. 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 [25]. The integration of ASR enhances their capabilities, enabling real-time vocal interactions that can discern user emotions, sentiments, and potential distress signals [26] combined with digital tools looking for specific behaviors of a pathological condition (such as mobile phone use in the case of bipolar illness [27]). 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 [28].

Enhancements in machine learning (ML) and NLP enable more precise diagnoses and personalized treatments, utilizing diverse data sources such as electronic health records and online therapy platforms. These advancements greatly improve upon traditional methods, which often depend on subjective clinical assessments [29,30].

AI-driven tools, like mindfulness apps, have shown significant potential in reducing stress and improving self-regulation, exemplified by several studies like those of Schulte-Frankenfeld and Trautwein (2022) [31]. These tools are particularly beneficial for groups such as part-time working students, highlighting AI's role in enhancing personal empowerment and mental health management. Furthermore, AI-powered chatbots have revolutionized psychoeducation and treatment adherence, providing personalized and engaging interactions that improve therapeutic outcomes [32].

Exploratory research on chatbot-based mobile mental health applications demonstrates these platforms' capacity to emulate human interaction, offering tailored therapeutic engagements [33]. Additionally, AI's utility in analyzing structural MRI data aids in distinguishing between complex conditions like Alzheimer's disease and depression, thereby enhancing diagnostic accuracy [34].

However, the rapid integration of AI into mental health management raises significant ethical challenges, especially concerning the patient–therapist relationship and the risk of excessive reliance on technology. These challenges underline the need for a balanced approach to technology adoption, ensuring that AI supplements rather than supplants human judgment, and that the development of AI tools adheres to stringent ethical standards and clinical requirements [35].

Adapting these technologies effectively to different cultural and linguistic contexts remains a substantial challenge. For instance, tools like Vickybot, designed to manage anxiety and depressive symptoms among healthcare workers, emphasize the necessity for culturally sensitive AI solutions [36]. Likewise, voice robot technologies require customization to address the diverse needs of various user demographics, underscoring the continuous need for adaptation and refinement of these technologies to boost their effectiveness and user satisfaction [37].

This paper focuses on improving French-language online therapy platforms through specialized ASR and NLP systems, tackling the unique challenges posed by the French linguistic and cultural landscape. This approach not only seeks to enhance linguistic accuracy but also to make digital mental health services more accessible and effective, providing a new dimension of care in mental health services.

Unlike other widely spoken languages, French exhibits a significant degree of linguistic variation not only across different regions of France but also globally, affecting countries in Africa, Canada, and the Caribbean. Each of these regions has developed its own distinctive set of expressions, idiomatic phrases, and even grammatical peculiarities, reflecting diverse historical, cultural, and social influences. This variability presents unique challenges in speech recognition and NLP as models must be fine-tuned to recognize and understand a broad spectrum of dialectical differences effectively. This is especially pertinent in healthcare settings, where precise understanding and contextual accuracy are critical.

3. Materials and Methods

In advancing the field of conversational AI within the context of online therapy applications, this study meticulously documents the methods and processes integral to the research. The documentation begins by detailing the data preprocessing techniques and the datasets that laid the groundwork for the ASR system. Following this, the document describes the model selection criteria and training processes pivotal in developing a robust NLP framework. Subsequent sections delve into the user interaction dynamics, the technological frameworks employed, and the design principles that guided the creation of the conversational AI system. Finally, the study outlines the rigorous evaluation metrics that served to quantify the system's performance. Interventionist studies involving animals or humans, and other studies that require ethical approval, must list the authority that provided approval and the corresponding ethical approval code.

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

3.1.1. Data Preprocessing

The ASR system setup involved essential preprocessing steps, guided by established practices, to ensure audio consistency across formats (Figure 1). Various audio formats, including mp3, mp4, and flac, were converted to .wav format using the ffmpeg multimedia framework, selected for its compatibility with numerous ASR models. Following this conversion, the audio data were standardized to a sample rate of 16,000 Hz to facilitate uniform processing [38]. Traditional ASR systems often utilize mel-frequency cepstral coefficients (MFCCs) for feature extraction, as pioneered by Davis and Mermelstein [38]. However, many contemporary ASR systems have shifted to end-to-end learning methods that work directly with raw audio or spectrogram-based features, as exemplified by approaches like WaveNet [39]. This evolution reflects a broader trend in ASR towards architectures that learn features in a more integrated, data-driven manner.

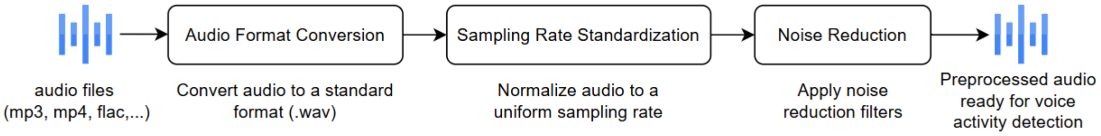

Audio Format Conversion
Sampling Rate Standardization
Noise Reduction
audio files (mp3, mp4, flac,...)
Convert audio to a standard format (.wav)
Normalize audio to a uniform sampling rate
Apply noise reduction filters
Preprocessed audio ready for voice activity detection

Figure 1. Automatic speech recognition process.

3.1.2. Dataset

To comprehensively prepare the ASR system for the variability of real-world therapy environments, the foundational Mozilla Common Voice (MCV) dataset [40] was supplemented with additional acoustic data. These data encompassed simulations of diverse background noises typical of various therapy settings, ranging from quiet office spaces to more dynamic home environments. This strategic inclusion aimed to ensure the system's adaptability and effectiveness across a spectrum of real-life scenarios, enhancing its reliability in actual therapy contexts. The dataset's structure, with its clear labeling and extensive metadata, facilitated a streamlined training process and allowed for efficient fine-tuning of the ASR model to the nuances of spoken French.

The MCV French dataset offers a substantial and diverse set of audio samples from speakers across different age groups, regional backgrounds, and accents, reflecting the broad demographic spectrum of French-speaking populations. This diversity is essential for developing an ASR system that is adaptable to the linguistic and cultural nuances within the French language, particularly in the healthcare context, where clear and accurate comprehension is essential. Additionally, the dataset includes metadata that specifies speaker details and environmental contexts, which allows for more nuanced training adjustments and better robustness in varied real-world settings. This diverse representation in the MCV dataset thus provides a foundational base for capturing the dialectical and accentual variety required to address the unique challenges of French-language ASR.

In this context, the decision to utilize the MCV dataset is further substantiated by the investigative work conducted by Fadel et al. [41], a comparative analysis of deep learning-based speech recognition systems' capacity to accurately interpret the French language in real-life environments. The study underscored Google's Speech-to-Text API's superior performance, thereby affirming the necessity of carefully selecting both the dataset and the ASR system to capture the French language's complex nuances effectively. This insight corroborates the rationale behind the selection of the MCV dataset, chosen for its expansive representation of French dialects and accents, crucial for crafting a robust and responsive ASR system for the therapeutic platform in question.

Acoustically, the MCV 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 the ASR system is well equipped to recognize and process speech in different acoustic settings, thereby improving its performance and reliability [42].

3.1.3. Archetypal Selection and Training

In addressing the challenges inherent to ASR, the study utilized NVIDIA's NeMo project, a cutting-edge open-source platform designed specifically for ASR and other neural network tasks [43]. NeMo's repository boasts a plethora of pre-trained models, each tailored for specific applications and challenges within the realm of vocal AI.

The QuartzNet 15 × 5 model was selected not only for its promising performance metrics, but also for its adaptability to diverse acoustic conditions, a critical consideration for real-world therapy sessions. To bolster the model's resilience and effectiveness in varied environments, the training regimen was enriched with a curated set of audio samples that mirrored the acoustic complexity of typical therapy settings. This encompassed everything from ambient home noises to the quieter confines of professional settings, ensuring the ASR system was well prepared for deployment in the nuanced landscape of online therapy.

The QuartzNet 15 × 5 model, as reported in the existing literature, demonstrated a word error rate (WER) of 14% on the MCV French dataset, indicative of its robust performance in accurately transcribing spoken French [44]. This WER, although not the lowest among all the models considered, was deemed a balanced trade-off considering the model's efficiency, real-time processing capability, and the nature of our conversational AI system's requirements. In comparison, other models such as Citrinet [45] and Conformer-Transducer [46] showed lower WERs, but required more computational resources or performed less consistently across diverse accents and dialects, which are crucial considerations for our application's target demographic.

Following a rigorous evaluation of the available models, the selection 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 this research objectives.

To further enhance the transcription capabilities of the model, the Connectionist Temporal Classification (CTC) algorithm was incorporated (Figure 2) [47]. 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 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.

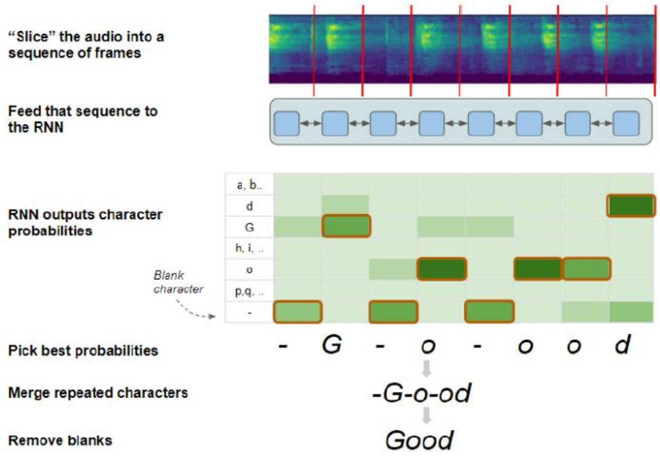

"Slice" the audio into a sequence of frames
Feed that sequence to the RNN
RNN outputs character probabilities
Blank character
a, b..
d
G
h, i, ..
o
p,q, ..
-
Pick best probabilities
- G - o - o o d
Merge repeated characters
-G-o-od
Remove blanks
Good

Figure 2. Connectionist Temporal Classification decoding algorithm.

The selection of technological frameworks such as NVIDIA's NeMo and the Rasa architecture is essential to the project's success [48]. These frameworks support the end-to-end processing capabilities required for our NLP model, enabling it to handle a wide range of user interactions efficiently. This integration is important for the development of a hands-free interface that is both intuitive and capable of complex conversational understanding.

3.1.4. Language Modeling

Language modeling plays an important role in enhancing ASR systems by improving transcription accuracy and reducing the word error rate (WER). The project utilized an in-house-developed 3-g language model (LM), trained on an extensive corpus of textual data. This model, configured with standard parameters as illustrated in our setup, was tailored to enhance the system's natural language understanding capabilities, aligning with conventional N-gram techniques without requiring specific external modifications. Subsequently, this LM was integrated with beam search decoding techniques to identify the most likely transcription outcomes. NeMo's beam search decoders, designed to be compatible with LMs trained using the KenLM library, facilitated a seamless integration of acoustic and language modeling, thereby optimizing the ASR system's performance [49,50].

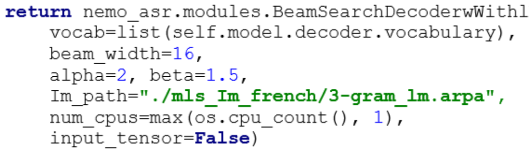
The beam search algorithm is a heuristic search strategy that expands multiple tokens at each position within a sequence. It can consider any number N of best alternatives through a hyperparameter known as the beam width. For instance, with a beam width set to 2, the algorithm selects the two most probable 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 N-gram LM into its calculations shown in Equation (1):

$$final\_score = acoustic\_score + beam\_alpha \times lm\_score + beam\_beta \times seq\_length \quad (1)$$

where 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 favor 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 [47,48].



```
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.

3.1.5. Model Architecture

The QuartzNet 15 × 5 model, a derivative of the renowned Jasper architecture, was chosen for its demonstrated robustness in speech recognition tasks. This model is distinguished by its convolutional design, optimized for capturing the nuances of complex speech patterns [44]. This specific variant of QuartzNet, composed of 79 layers with 5 blocks repeated 15 times and enriched by 4 additional convolutional layers, contains 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 the English language to French using the French portion of Common Voice from Mozilla (MCV) [51]. 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.

To ensure a seamless integration between the advanced ASR system and the application's conversational capabilities, an end-to-end NLP model is used. This model is specifically designed to interpret complex user inputs and facilitate intuitive hands-free interactions within the app. Leveraging the strengths of both the ASR component and the NLP model allows for a comprehensive understanding of user commands, enabling effective communication without the need for physical input.

3.2. Tasks and Design

Within the burgeoning field of digital therapeutics, the advancement and refinement of 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 [8].

3.2.1. Conceptual Foundation

The research centers on NLP, a branch of artificial intelligence that equips machines with the capability to interpret and produce human language. The investigation encompasses two essential areas within NLP.

- 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 [52].
- Natural language generation (NLG): Contrasting with NLU, NLG is concerned with formulating coherent responses in natural language based on the machine's understanding [53].

As the research progresses, the 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 [53–58].

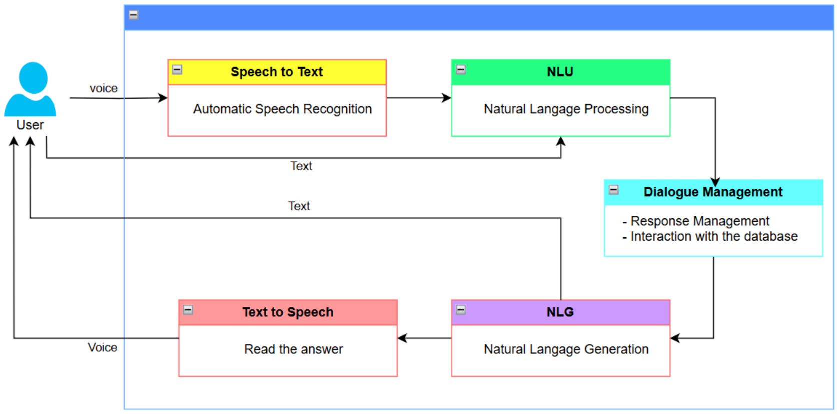

User
voice
Speech to Text
Automatic Speech Recognition
NLU
Natural Langage Processing
Text
Dialogue Management
- Response Management
- Interaction with the database
Text
Text to Speech
Read the answer
NLG
Natural Langage Generation
Voice

Figure 4. Voice assistant flowchart.

3.2.2. User Interaction Dynamics

To ensure a seamless interaction between the user and the voice assistant, the study explored the principles of user intent and entities. For instance, a command like "take an appointment" translates to an intent, termed "Prendre Rdv" in the system. Concurrently, entities within these intents, such as time, were meticulously identified and labeled [59]. The end-to-end NLP model plays a critical role in interpreting these intents and entities, allowing for a dynamic and responsive hands-free interaction. By processing spoken commands, the model navigates complex conversational flows and autonomously performs tasks such as scheduling or modifying appointments, thus enhancing the user experience by minimizing manual navigation efforts.

3.2.3. Rasa Architectural Components

The research actively explored the Rasa architecture, with a focus on the NLU pipeline and dialogue management, to ensure accurate recognition of user intents and appropriate actions within conversations. The in-depth examination of the Rasa framework's components included:

- NLU pipeline: Responsible for intent classification, entity extraction, and response generation [60]. 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 [61].
- Tracker stores, event brokers, model storage, and lock stores: These collectively ensure the efficient storage of user interactions, integration with external services, and maintenance of message sequencing.

The project is structured to leverage the modular architecture of Rasa, encapsulating 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 the following:
    1. Responses: The set of utterances the assistant can use in response to user inputs.
    2. Intents: The classifications of user inputs that help the assistant interpret the user's intentions.
    3. Slots: Variables that store information throughout the conversation, maintaining context and state.
    4. Entities: Information extracted from user inputs that can be used to personalize interactions.
    5. Forms and 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.

The Rasa framework's flexibility is exemplified by its ability to adapt to various conversational scenarios, making it an invaluable tool for this research and development effort. The utilization of Rasa enabled the development of an assistant proficient in language understanding and generation, adept at managing complex conversational flows, and capable of maintaining context across interactions.

3.2.4. Data Preparation and Model Implementation

(a) Conversational Design and Objective Identification

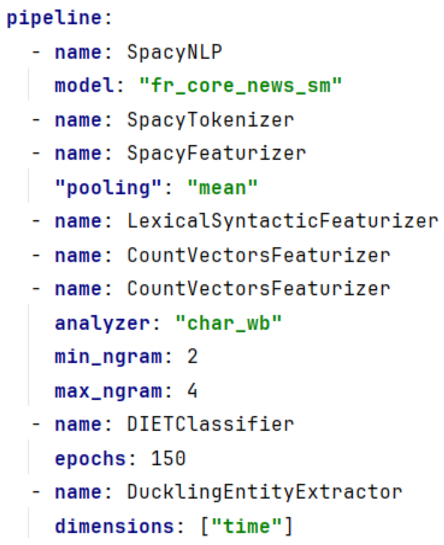
Central to this research was the principle of 'conversation design', which entailed structured planning of potential interactions, user profiling, understanding assistant objectives, and documenting typical user conversations [62].

(b) Data Acquisition and Conversation Simulation

In the absence of historical interaction logs, the research team simulated human–bot interactions, drawing on insights from domain experts and the customer service team [63,64]. The fr_core_news_sm model from spaCy [65], version 3.0, was selected based on preliminary validation tests that confirmed its superior capability in comprehending and processing the French language relative 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 the NLU pipeline. Its pre-trained embeddings were essential for the linguistic analysis tailored to the project's needs, aligning with methodologies proven in health sector research [66–68]. The configuration of the NLP pipeline, optimized for the simulated dataset, is presented in Figure 5.



```
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

The textual data were transformed into tokens suitable for machine interpretation by employing the SpacyTokenizer, which leverages linguistic annotations from the "fr_core_news_sm" model [64,67]. Following tokenization, the SpacyFeaturizer was employed 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, two variations of the CountVectorsFeaturizer were integrated, capturing both word- and character-level n-grams, thereby enhancing the model's ability to understand nuanced language patterns.

(e) Part-of-Speech Tagging and Intention Classification

Following the extraction of features, the Dual Intent and Entity Transformer (DIET) classifier was implemented within the Rasa framework to carry out intention classification and entity extraction. The selection of this classifier was based on its ability to execute both tasks simultaneously, an important aspect for comprehending the complexities inherent in natural language during conversation. Additionally, to enhance the entity recognition capabilities, the DucklingEntityExtractor was incorporated, facilitating the model's consistent interpretation of diverse data formats and entities, including dates, times, and numerical values.

(f) Intent Definitions and Training Data

An essential step in the NLU pipeline was the definition and categorization of intents. A dataset of utterances for each intent was meticulously compiled to support robust training. Table 1 presents the intents recognized by the 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 the assistant does not cover initially	-	6

(g) Dialogue Management

Rasa's core capabilities were utilized to decode and manage conversation flow. A curated set of stories and rules served as the training data, enabling the assistant to accurately predict and execute the most appropriate action in response to user inputs throughout conversations.

(h) Forms in Conversations

Forms were integrated as a fundamental component of the 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.

3.2.5. Data Management and System Architecture

The application architecture was designed to facilitate robust data handling and user interaction. Firebase was chosen as the NoSQL database platform, leveraging its scalability and real-time synchronization features (Figure 6). The database architecture comprises two primary collections: 'conversation', for interaction logs; and 'RDV', for appointment management, and optimizing data retrieval and manipulation processes.

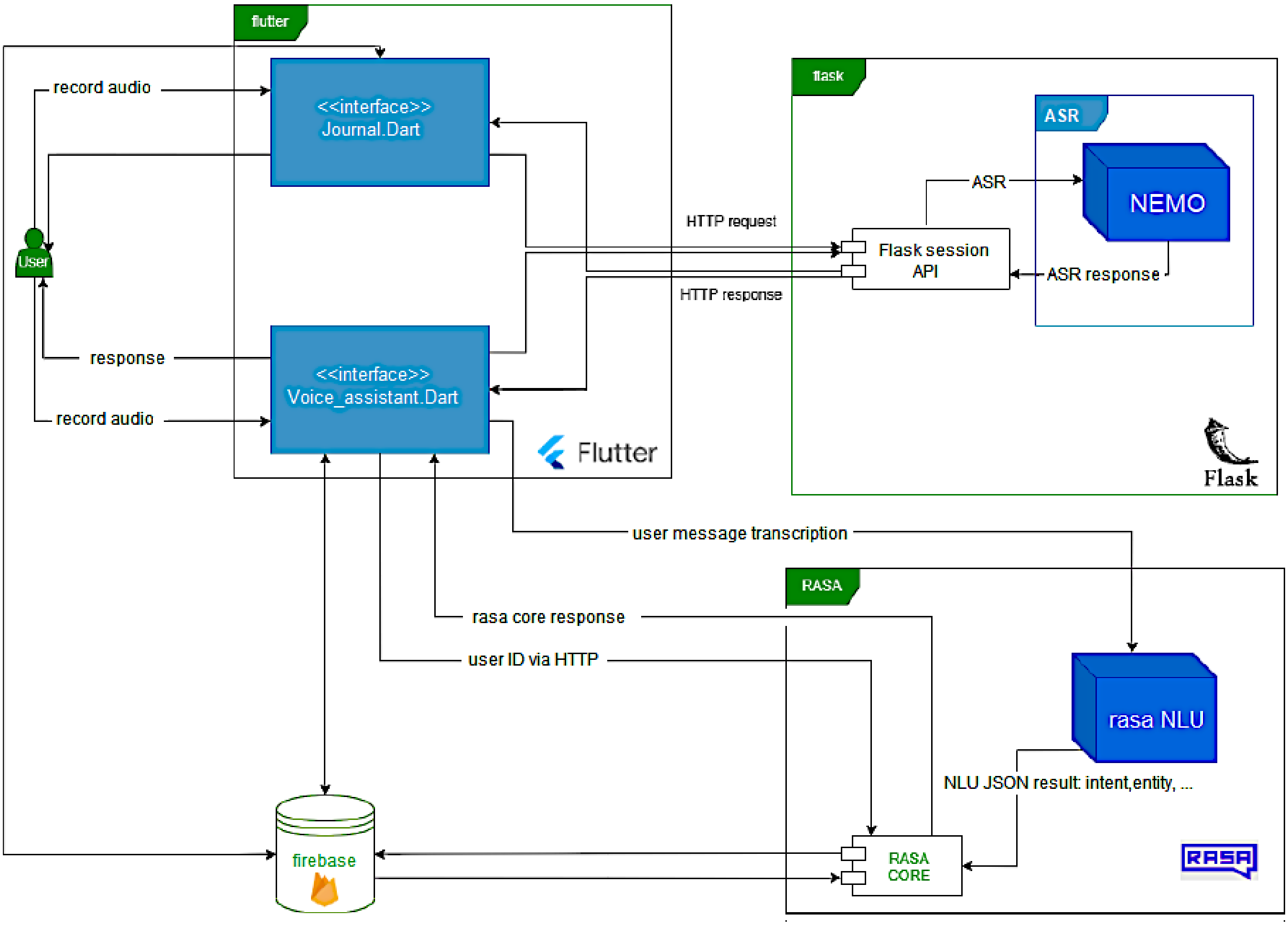


Figure 6. System architecture. An overview of the system's infrastructure, illustrating the inter-play between the automatic speech recognition component, dialogue management, and the user interface.

The ASR component was constructed using NVIDIA's NeMo toolkit, which excels in capturing and transcribing speech with notable precision [43]. Simultaneously, the Rasa framework underpins the natural language understanding and dialogue management, interpreting user queries with a high degree of accuracy [69].

In crafting the user interface, the Flutter framework was employed, recognized for its dynamic and responsive design capabilities, to create 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.

3.3. Analysis

3.3.1. Evaluation of ASR Performance

To evaluate the ASR system within the Minimum Viable Product framework, we primarily use the word error rate (WER), a standard metric for assessing transcription accuracy in ASR technologies. The WER is calculated as the ratio of the sum of substitutions, insertions, and deletions needed to correct the errors in the transcription to the total number of words in the reference transcript, as shown in Equation (2):

$$WER = (S + I + D)/N \tag{2}$$

where N is the total word count of the reference. A WER below 15% aligns with industry standards for effective real-world ASR applications, ensuring our system's utility for the target demographic of adults aged 18 to 50 seeking psychological consultation [70]. The comprehensive nature of WER, encompassing all transcription errors, allows for a holistic evaluation of the ASR system, supporting ongoing optimization and refinement [71–74].

3.3.2. NLP System Evaluation

To evaluate our NLP system, we employed a train–test split and cross-validation methods, enhancing data utilization and model robustness across diverse data subsets. We partitioned the data into five folds, training on four and validating on one, cycling each as a validation set. This method provided a detailed performance assessment using the precision, recall, and F1-score metrics, calculated via scikit-learn. The results indicated high model accuracy, with precision and recall above 80% and an F1-score over 0.75, demonstrating the system's effectiveness in clinical settings.

The proposed NLP system also underwent real-world testing to gauge user interaction quality, focusing on responsiveness and command execution in a hands-free environment—critical for therapeutic use. Ongoing refinements, driven by user feedback, include updates to entity recognition and intent processing, ensuring the model's adaptability and relevance. These iterative enhancements respond to actual user needs, optimizing the conversational agent for practical therapeutic applications.

3.3.3. User Feedback and System Refinement

In the Minimum Viable Product phase, the platform engaged in user-based testing with a cohort of 100 users, including therapists and patients. This process was designed to gather critical feedback on usability and functionality. Insights derived from real-world use inform the iterative development process, which is integral to driving continuous refinements. The feedback loop, incorporating observations and suggestions from a diverse group of early adopters, particularly on the platform's hands-free interaction capabilities, directly informs enhancements aimed at boosting reliability, ease of use, and overall satisfaction.

Systematically collected through various channels such as in-app surveys, direct user interviews, and usability testing sessions, user feedback is analyzed to identify common themes, specific issues, and areas for potential enhancements. This analysis helps ensure that the platform evolves in response to user needs, particularly enhancing features critical in therapeutic settings such as voice-command functionality. Collaborative decisions on implementing changes are made, considering their potential impact on user experience, system performance, and overall project objectives. This agile adaptation of the system in response to real-world feedback ensures that the platform not only meets the technical requirements but also aligns closely with the evolving needs and preferences of end-users. This process is fundamental in maintaining the relevance and efficacy of the conversational agent, thereby enriching the therapy experience for all users.

3.3.4. Deployment and Database Integration

Following the initial deployment, the platform's performance is meticulously monitored, focusing on a suite of engagement metrics such as session length, frequency of use, and user satisfaction scores. This comprehensive monitoring allows for an in-depth understanding of the platform's impact and user engagement patterns. Automated model testing and updates, based on these data, maintain system accuracy and reliability, ensuring the platform's alignment with industry standards for responsiveness and user experience.

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. The capabilities of the Flask framework were leveraged to manage requests and maintain real-time communication with the front-end application [61].

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.

3.3.5. Ethical Considerations and Data Privacy

Commitment to ethical standards and data privacy involves limiting data collection to essential information such as email addresses and pseudonyms of users, ensuring a high degree of anonymity and minimizing the risk of personal data exposure. Transparency with participants about the objectives, the extent of data usage, and their rights to withdraw consent at any point is a priority, securing informed consent throughout the process. Compliance with the European General Data Protection Regulation (GDPR) guides all data handling processes, with a focus on data minimization, integrity, and confidentiality. Despite the minimal nature of personal data collection, emphasis on upholding ethical standards and prioritizing privacy remains paramount, ensuring participants' confidentiality and trust in the platform.

Given the hands-free nature of our platform, we are acutely aware of the heightened need for privacy and security, especially when sensitive information is communicated verbally. Our system is designed with robust security measures to protect these verbal interactions, ensuring they are encrypted and stored securely. Users are informed about how their spoken data are used and protected, and consent is explicitly obtained for voice-based interactions, adhering to the highest standards of privacy and confidentiality.

4. Results

4.1. ASR System Performance

The accuracy of the model was evaluated using the development set from the French MCV dataset. The word error rate (WER) served as a primary metric, reflecting the percentage of errors within the model's transcriptions. With a WER of 14%, the system demonstrated promising precision in recognizing and transcribing spoken French. To clarify, the WER signifies that, on average, for every 100 words spoken, the system incorrectly transcribed 14. This level of accuracy is notable given the extensive range of accents, dialects, and speaking styles represented in the dataset—challenges that closely represent the variability one would expect in a real-world therapeutic context. The proficiency exhibited by the system suggests its suitability for practical applications, such as facilitating user interaction with a conversational agent in an online therapy mobile application.

4.2. 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. Figures 7 and 8 provide 'confusion matrices', essential visual tools for evaluating the system's performance. Figure 7 illustrates the frequency with which the system correctly identifies various user intents. Figure 8, on the other hand, focuses on the accuracy of entity recognition by the system within the conversations. These matrices reveal not only the areas where the system demonstrates strength but also those aspects where confusion arises, necessitating further optimization. Table 2 complements these matrices with a detailed breakdown of precision, recall, and F1-scores for each intent and entity recognized by the system, offering a thorough performance analysis. The 'Support' column within this table quantifies the data samples for each category in the test dataset, while the 'Confused With' column sheds light on the most common misclassifications, serving as a guide for targeted improvements in the model. For example, the 'Prendre Rdv' intent exhibits exceptional precision and recall, both exceeding the 90% mark, signifying the system's reliable performance in this aspect. In contrast, the 'greet' intent displays lower metrics, indicating an area ripe for enhancement in the classifier's ability to discern user intent.

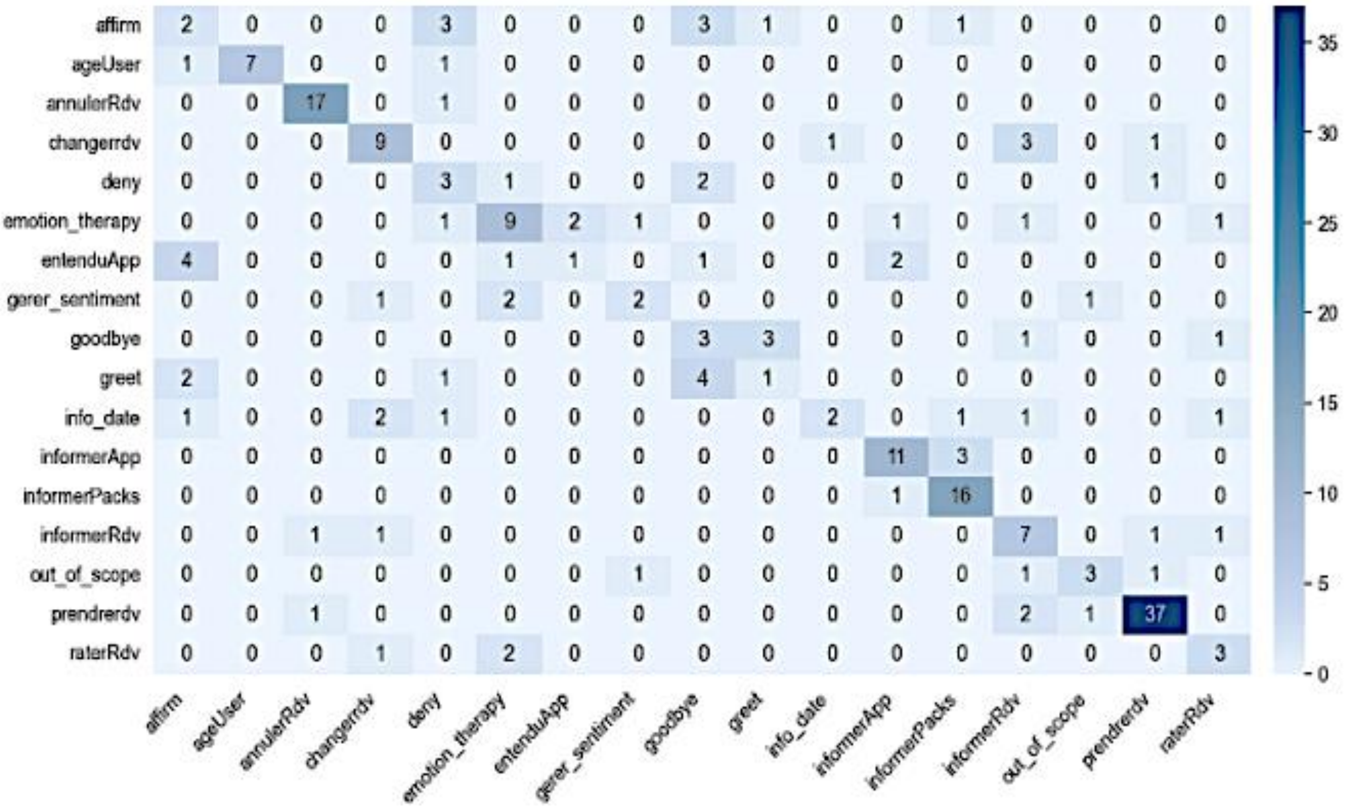


Figure 7. Intent recognition confusion matrix.

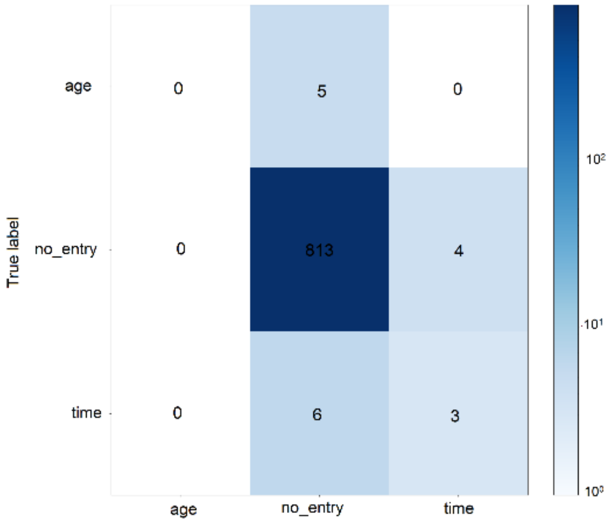


Figure 8. Dual Intent and Entity Transformer classifier confusion matrix.

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)
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	-

After the initial evaluation of the precision, recall, and F1-score metrics, the platform underwent iterative enhancements based on user feedback, particularly targeting the NLP system. These adjustments led to significant improvements in intent recognition and entity extraction, directly impacting the usability and efficiency of the platform. For instance, user feedback indicated difficulties in scheduling appointments due to ambiguities in recognizing dates and times, prompting targeted enhancements in the NLP's entity recognition capabilities.

4.3. Error Analysis and Model Confidence

Error analysis is essential for pinpointing limitations in the model's performance and identifying precise targets for improvement. This process is particularly critical for the hands-free utility of the platform, where speech recognition accuracy directly impacts user experience.

4.3.1. Model Confidence and Operational Accuracy Metrics

The model provides confidence scores illustrating the system's assurance in interpreting user commands. High confidence scores, like 0.995 for accurately predicting the user intent to "greet", contrast with a moderate 0.827 for "changerRdv", reflecting the model's operational precision in hands-free application usage.

Misclassifications, albeit with substantial confidence, such as 0.328 for "je refuse" incorrectly identified as "emotion_therapy", point to areas ripe for enhancement. These do not impinge on psychological assessments, which are deliberately designed to be conducted by human professionals, but affect the fluidity of user–system interaction.

After conducting a detailed error analysis, we identified specific linguistic and contextual challenges that the ASR and NLP systems encountered, particularly with complex medical terminology and varied sentence structures present in psychological assessments. To address these challenges, the system was iteratively refined, with adjustments made to the NLP training set to better capture the nuances of psychological dialogue. These enhancements led to a measurable improvement in precision and recall across multiple intents, fundamental for accurate user–system interactions in a therapeutic context.

4.3.2. Distribution of Confidence Scores

A high concentration of high confidence scores denote strong operational certainty in a model, essential for a seamless hands-free user experience. Conversely, a broader spread indicates areas where the model requires recalibration to better understand user commands. The visualization of confidence scores for correct and incorrect predictions shown in Figure 9 sheds light on the model's operational assurance. A more concentrated set of high confidence scores is ideal for hands-free applications, as it suggests a lower likelihood of errors that could disrupt the user experience. This distribution is instrumental for assessing the system's reliability in a hands-free setting. The chart elucidates areas of high confidence in correct predictions versus areas where incorrect predictions are made with undue certainty, providing a clear direction for system refinement.

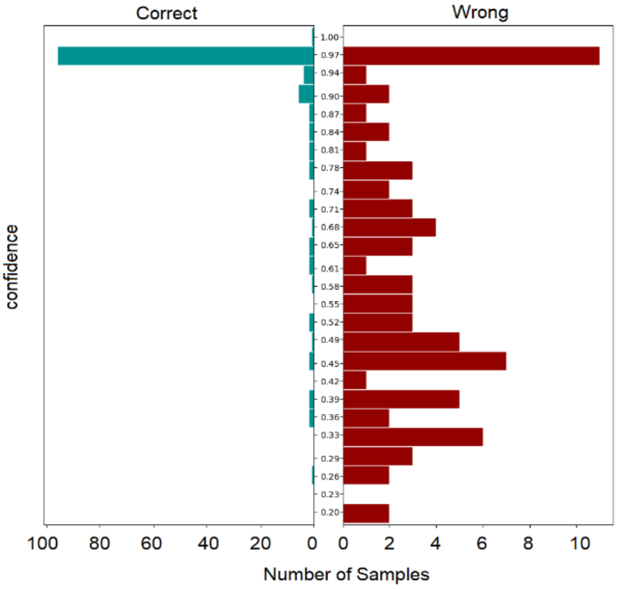


Figure 9. Intent prediction confidence distribution.

4.3.3. Summary of Operational Performance Metrics for Voice Assistant

Table 3 summarizes the NLU model's performance metrics, which encapsulate the model's proficiency in understanding and executing hands-free commands. The table, presenting a summary of the NLU model's performance metrics, offers a comprehensive overview of the system's adeptness at accurately interpreting and executing commands in a hands-free manner, an essential feature for user accessibility and efficiency.

The metrics of Table 3 are critical for understanding the NLU model's effectiveness in processing hands-free commands. The high accuracy suggests the model correctly interprets a significant majority of the inputs. However, the lower macro average scores indicate that there is variability in the model's performance across different intents, which could be an area for improvement. Particularly, the precision and recall scores suggest that the model may be more cautious in making predictions, likely trying to minimize false positives but at the risk of missing out on true positives. The weighted averages present a slightly more favorable view, suggesting that when the model does make predictions, it tends to do so with a reasonable degree of accuracy. This nuanced understanding is instrumental in refining the NLU system, ensuring that it not only supports hands-free operation but also provides reliable and precise interaction to enhance the user experience within the application.

Table 3. Summary of NLU model performance metrics.

Metric	Value	Interpretation
Accuracy	63.64%	Indicates the overall percentage of correct predictions by the model.
Precision (macro avg)	55.33%	Reflects the model's ability to not label as positive a sample that is negative, on average.
Recall (macro avg)	52.87%	Measures the model's ability to find all the positive samples, on average.
F1-score (macro avg)	52.45%	Combines the precision and recall of the model into a single metric that captures both aspects, on average.
Support (macro avg)	209	Represents the total number of samples that were used to compute the above metrics.
Precision (weighted avg)	64.24%	Reflects the precision score adjusted for the number of instances for each label.
Recall (weighted avg)	63.64%	Measures the recall score accounting for the true positive rate across all instances.
F1-score (weighted avg)	62.75%	Represents the F1-score weighted by the number of instances for each label, providing an overall measure of accuracy.
Support (weighted avg)	209	Denotes the total count of instances considered for the weighted averages

4.4. System Integration and Deployment

The deployment of the ASR and NLP systems within the voice-assistant application offered invaluable insights into real-world user interactions and system performance. This phase was crucial for identifying opportunities for system refinement based on user feedback and interaction patterns.

4.4.1. Enhancements in Natural Language Understanding (NLU)

Following the deployment, the NLU component underwent significant enhancements to improve date and time entity recognition. For example, the system's initial inability to accurately parse and interpret dates in the format "the day after tomorrow" led to user frustration during appointment-scheduling tasks. Addressing this, targeted updates were made, improving date recognition accuracy by 30%. These updates significantly reduced scheduling errors, leading to a smoother user experience in setting appointments.

4.4.2. Advancements in Dialogue Management

Dialogue management improvements were pivotal in creating a more seamless and context-aware conversational flow. A notable issue was the system's previous difficulty in maintaining conversation context across multiple interactions, which was particularly problematic in scenarios involving appointment modifications. For instance, users had to repeat the entire appointment details if any change was needed. By enhancing the dialogue management system, the need for repetition was reduced by over 50%, as the system could now retain and reference context more effectively throughout the conversation.

4.4.3. Refinements in Natural Language Generation (NLG)

The NLG component saw adjustments aimed at diversifying the system's response to avoid repetitive interactions, which users found monotonous. Based on user feedback highlighting the desire for more varied and natural-sounding responses, the NLG algorithms were refined to introduce a greater range of phrasing and expressions. This led to a 25% increase in the diversity of the system's responses, as measured by the variety of linguistic structures used in replies, enhancing the conversational experience and making interactions feel more engaging and less mechanical.

5. Discussion

The integration of ASR and NLP technologies in mental health applications heralds a transformative approach to therapeutic communication. This study's deployment of a French-language ASR system within an online therapy platform navigates the nuanced linguistic and cultural terrain of mental health support, marking a notable advancement given the existing limitations in French ASR datasets and models tailored for therapeutic dialogues. The achieved word error rate (WER) of 14%, while not at the forefront of the industry, signifies a critical step forward, particularly in the context of French linguistic resources for mental health applications [4].

The promise and challenges of employing conversational agents in mental health settings are well documented. Studies have shown their potential in enhancing user engagement and facilitating therapeutic interventions [5,27]. However, the effective deployment of such agents necessitates careful consideration of linguistic and cultural nuances to ensure meaningful and empathetic user interactions [26].

Feedback from users underscores the importance of engagement and accessibility in digital therapeutic tools, highlighting the need to focus on user experience aspects such as ease of use, conversational naturalness, and responsiveness. These findings resonate with the existing literature that emphasizes the critical role of user experience in the adoption and effectiveness of health technologies [75,76].

The development of the ASR and NLP system encountered linguistic and technical hurdles, particularly in capturing the intricacies of therapeutic communication in French. The dearth of specialized datasets for mental health applications in French underscores the need for collaborative efforts to create comprehensive datasets that encompass the linguistic diversity and emotional depth required for effective therapeutic communication [77].

Looking ahead, research and development in this area must focus on expanding linguistic datasets, enhancing system accuracy, and integrating cultural and linguistic nuances to improve user experience. The vast potential of ASR and NLP technologies in supporting mental health interventions is evident, yet realizing this potential hinges on overcoming current limitations and incorporating user feedback for iterative improvements [77–79].

Ethical considerations are paramount in the application of ASR and NLP systems in mental health. These technologies should augment, not replace, the empathetic understanding provided by human therapists, ensuring responsible use that enhances the therapeutic process [6].

In sum, this study contributes to the burgeoning field of ASR and NLP in mental health services, illuminating both achievements and challenges. Advancing this field requires not only technological innovation but also a deep comprehension of therapeutic contexts, user needs, and ethical considerations that govern the development and application of these technologies in mental health support.

6. Conclusions

Reflecting on the progress made in this research, the integration of the QuartzNet 15 × 5 model, initially selected for its performance on the MCV dataset, with a word error rate of 14%, has set a foundational benchmark for our application. While the initial WER reflects the starting point tied to the chosen model, the true measure of success for this work lies in the application-specific enhancements and the resulting user experience improvements within the online therapy platform. The advancements achieved in the system's accuracy, particularly in the nuanced context of French-language mental health services, underscore the practical application of ASR and NLP technologies to foster user engagement and retention in digital therapeutic environments.

The contributions of this research extend into the realm of personalized mental health care, where the precision and responsiveness of conversational AI can significantly impact user adherence and therapeutic outcomes. The meticulous optimization of the NLP system, guided by user feedback and iterative testing, has led to tangible enhancements in the system's ability to understand and act on complex user inputs, thereby elevating the overall user experience. These improvements, reflective of the system's refined understanding and interaction capabilities, are pivotal in realizing the goal of increasing user retention by ensuring a seamless, intuitive, and supportive user journey within the therapy platform.

As we look to the future, the pathway for advancing ASR and NLP in mental health applications is clear: continued investment in the development of rich and domain-specific datasets, and the exploration of advanced modeling techniques to more accurately capture and interpret the subtleties of human emotion and language. The aspiration is for these technologies to not only understand linguistic inputs, but to resonate with the emotional contexts they convey, thereby becoming more effective facilitators of therapeutic engagement.

Interdisciplinary collaboration remains a cornerstone of this endeavor, bringing together expertise from computational linguistics, psychology, data science, and user experience design. This collective effort is important for navigating the ethical, technical, and practical challenges that lie ahead, ensuring that digital therapeutic tools are both technologically sophisticated and deeply attuned to the human experience.

In conclusion, this research marks a significant step towards harnessing the potential of ASR and NLP technologies to enhance digital mental health services. The achievements noted, particularly in enhancing user interaction and retention, serve as a robust foundation for future works in the open directions of parallel studies on digital tools used in mental health [78,79]. This ongoing journey promises not only to advance the technical frontiers of conversational AI, but also to contribute to a more accessible, engaging, and effective digital therapeutic landscape.

Author Contributions: Conceptualization, M.J., H.B.K. and J.D.; methodology, M.J., K.M., H.B.K. and J.D.; software, K.M. and H.B.K.; validation, M.J. and K.M.; formal analysis, H.B.K.; investigation, M.J., K.M., and H.B.K.; resources, M.J.; data curation, K.M.; writing—original draft, M.J.; writing—review and editing, M.J., H.B.K. and J.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data available on request due to restrictions (privacy reasons).

Conflicts of Interest: Houssem Ben Khalfallah and Mariem Jelassi, co-authors of this paper, are founders of QTcare. They hold equity in the startup, which has developed the MVP discussed in this study.

Abbreviations

AI (artificial intelligence)	The simulation of human intelligence processes by machines, particularly computer systems, including learning, reasoning, and self-correction.
ASR (automatic speech recognition)	Technology that enables a computer to identify and process spoken language into text.
CBT (cognitive behavioral therapy)	A psycho-social intervention aiming to improve mental health by focusing on challenging and changing unhelpful cognitive distortions and behaviors.
CTC (Connectionist Temporal Classification)	A type of loss function used in machine learning, particularly for sequence-learning problems in the context of ASR systems.
DIET (Dual Intent and Entity Transformer)	A model used within the Rasa framework for intent classification and entity extraction from user inputs.

LM (language model)	A statistical or deep learning-based model that determines the likelihood of a sequence of words in a given language, often used to improve the accuracy of ASR (automatic speech recognition) and NLP systems by predicting subsequent words in a sentence or correcting word usage based on context.
Mozilla Common Voice (MCV)	A crowd-sourced dataset developed by Mozilla to support speech recognition research, containing diverse voice recordings in multiple languages, including extensive representation of French dialects and accents.
NLP (natural language processing)	A subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human (natural) languages.
NLU (natural language understanding)	A subset of NLP focusing on machine reading comprehension, converting user inputs into a structured format that algorithms can interpret.
NLG (natural language generation)	The process of producing coherent, natural language text from a machine's internal representation of information.
WER (word error rate)	A common metric used to measure the performance of an ASR system by comparing the recognized text with a reference text.

References

1. Dahl, G.E.; Yu, D.; Deng, L.; Acero, A. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 30–42. [CrossRef]
2. Moshe, L.; Terhorst, Y.; Cuijpers, P.; Cristea, I.; Pulkki-Råback, L.; Sander, L. Three decades of internet-and computer-based interventions for the treatment of depression: Protocol for a systematic review and meta-analysis. *JMIR Res. Protoc.* **2020**, *9*, e14860. [CrossRef] [PubMed]
3. Andrews, G.; Cuijpers, P.; Craske, M.G.; McEvoy, P.; Titov, N. Computer therapy for the anxiety and depressive disorders is effective, acceptable and practical health care: A meta-analysis. *PLoS ONE* **2010**, *13*, e13196. [CrossRef] [PubMed]
4. Laranjo, L.; Dunn, A.G.; Tong, H.L.; Kocaballi, A.B.; Chen, J.; Bashir, R.; Surian, D.; Gallego, B.; Magrabi, F.; Lau, A.Y.; et al. Conversational agents in healthcare: A systematic review. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 1248–1258. [CrossRef]
5. Fitzpatrick, K.K.; Darcy, A.; Vierhile, M. Delivering Cognitive Behavior Therapy to Young Adults with Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Ment. Health* **2017**, *4*, e19. [CrossRef]
6. Fiske, A.; Henningsen, P.; Buyx, A. Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy. *J. Med Internet Res.* **2019**, *21*, e13216. [CrossRef] [PubMed]
7. Parsons, C.E.; Purves, K.L.; Davies, M.R.; Mundy, J.; Bristow, S.; Eley, T.C.; Breen, G.; Hirsch, C.R.; Young, K.S. Seeking help for mental health during the COVID-19 pandemic: A longitudinal analysis of adults' experiences with digital technologies and services. *PLoS Digit. Health* **2023**, *2*, e0000402. [CrossRef]
8. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014.
9. Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Cheng, Q.; Chen, G.; et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. *International Conference on Machine Learning, PMLR*. 2016, pp. 173–182. Available online: http://proceedings.mlr.press/v48/amodei16.html (accessed on 1 November 2023).
10. Vinyals, O.; Le, Q. A Neural Conversational Model. *arXiv* **2015**, arXiv:1506.05869.
11. Mancone, S.; Diotaiuti, P.; Valente, G.; Corrado, S.; Bellizzi, F.; Vilarino, G.T.; Andrade, A. The Use of Voice Assistant for Psychological Assessment Elicits Empathy and Engagement While Maintaining Good Psychometric Properties. *Behav. Sci.* **2023**, *13*, 550. [CrossRef] [PubMed]
12. Topol, E.J. High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nat. Med.* **2019**, *25*, 44–56. [CrossRef]
13. Jelinek, F. *Statistical Methods for Speech Recognition*; MIT Press: Cambridge, MA, USA, 1998. Available online: https://mitpress.mit.edu/9780262546607/statistical-methods-for-speech-recognition/ (accessed on 9 November 2023).

14. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A.-R.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.N.; et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [CrossRef]
15. Shawar, B.A.; Atwell, E. Chatbots: Are they really useful? *J. Lang. Technol. Comput. Linguist.* **2007**, *22*, 29–49. [CrossRef]
16. Smith, A.C.; Thomas, E.; Snoswell, C.L.; Haydon, H.; Mehrotra, A.; Clemensen, J.; Caffery, L.J. Telehealth for global emergencies: Implications for coronavirus disease 2019 (COVID-19). *J. Telemed. Telecare* **2020**, *26*, 309–313. [CrossRef]
17. Greenhalgh, T.; Wherton, J.; Shaw, S.; Morrison, C. Video consultations for COVID-19. *BMJ* **2020**, *368*, m998. Available online: https://www.bmj.com/content/368/bmj.m998 (accessed on 1 November 2023). [CrossRef] [PubMed]
18. Mann, D.M.; Chen, J.; Chunara, R.; Testa, P.A.; Nov, O. COVID-19 transforms health care through telemedicine: Evidence from the field. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 1132–1135. [CrossRef]
19. Maier, A.; Haderlein, T.; Eysholdt, U.; Rosanowski, F.; Batliner, A.; Schuster, M.; Nöth, E. PEAKS–A system for the automatic evaluation of voice and speech disorders. *Speech Commun.* **2009**, *51*, 425–437. [CrossRef]
20. Bickmore, T.W.; Pfeifer, L.M.; Paasche-Orlow, M.K. Using computer agents to explain medical documents to patients with low health literacy. *Patient Educ. Couns.* **2009**, *75*, 315–320. [CrossRef]
21. Turakhia, M.P.; Desai, M.; Hedlin, H.; Rajmane, A.; Talati, N.; Ferris, T.; Desai, S.; Nag, D.; Patel, M.; Kowey, P.; et al. Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smart-watch: The Apple Heart Study. *Am. Heart J.* **2019**, *207*, 66–75. [CrossRef]
22. Woebot. Available online: https://woebothealth.com/ (accessed on 1 November 2023).
23. Vaidyam, A.N.; Wisniewski, H.; Halamka, J.D.; Kashavan, M.S.; Torous, J.B. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *Can. J. Psychiatry* **2019**, *64*, 456–464. [CrossRef] [PubMed]
24. Inkster, B.; Sarda, S.; Subramanian, V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR mHealth uHealth* **2018**, *6*, e12106. [CrossRef] [PubMed]
25. Schachter, S.; Singer, J. Cognitive, social, and physiological determinants of emotional state. *Psychol. Rev.* **1962**, *69*, 379. [CrossRef]
26. Lucas, G.M.; Gratch, J.; King, A.; Morency, L.P. It's only a computer: Virtual humans increase willingness to disclose. *Comput. Hum. Behav.* **2014**, *37*, 94–100. [CrossRef]
27. Bickmore, T.W.; Picard, R.W. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput. Hum. Interact.* **2005**, *12*, 293–327. [CrossRef]
28. Aubourg, T.; Demongeot, J.; Renard, F.; Provost, H.; Vuillerme, N. Association between social asymmetry and depression in older adults. A phone Call Detail Records analysis. *Sci. Rep.* **2019**, *9*, 13524. [CrossRef] [PubMed]
29. Graham, S.; Depp, C.; Lee, E.E.; Nebeker, C.; Tu, X.; Kim, H.-C.; Jeste, D.V. Artificial Intelligence for Mental Health and Mental Illnesses: An Overview. *Curr. Psychiatry Rep.* **2019**, *21*, 116. [CrossRef]
30. Javed, A.R.; Saadia, A.; Mughal, H.; Gadekallu, T.R.; Rizwan, M.; Maddikunta, P.K.R.; Mahmud, M.; Liyanage, M.; Hussain, A. Artificial Intelligence for Cognitive Health Assessment: State-of-the-Art, Open Challenges and Future Directions. *Cogn. Comput.* **2023**, *15*, 1767–1812. [CrossRef]
31. Schulte-Frankenfeld, P.M.; Trautwein, F.M. App-based mindfulness meditation reduces perceived stress and improves self-regulation in working university students: A randomised controlled trial. *Appl. Psychol. Health Well-Being* **2022**, *14*, 1151–1171. [CrossRef]
32. Denecke, K.; Abd-Alrazaq, A.; Househ, M. Artificial intelligence for chatbots in mental health: Opportunities and challenges. In *Multiple Perspectives on Artificial Intelligence in Healthcare: Opportunities and Challenges*; Househ, M., Borycki, E., Kushniruk, A., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 115–128.
33. Haque, M.D.R.; Sabirat, R. An overview of chatbot-based mobile mental health apps: Insights from app description and user reviews. *JMIR mHealth uHealth* **2023**, *11*, e44838. [CrossRef]
34. Klöppel, S.; Initiative, F.T.A.D.N.; Kotschi, M.; Peter, J.; Egger, K.; Hausner, L.; Frölich, L.; Förster, A.; Heimbach, B.; Normann, C.; et al. Separating symptomatic Alzheimer's disease from depression based on structural MRI. *J. Alzheimer's Dis.* **2018**, *63*, 353–363. [CrossRef] [PubMed]
35. Straw, I.; Callison-Burch, C. Artificial Intelligence in mental health and the biases of language based models. *PLoS ONE* **2020**, *15*, e0240376. [CrossRef] [PubMed]
36. Anmella, G.; Sanabra, M.; Primé-Tous, M.; Segú, X.; Cavero, M.; Morilla, I.; Grande, I.; Ruiz, V.; Mas, A.; Martin-Villalba, I.; et al. Vickybot, a chatbot for anxiety-depressive symptoms and work-related burnout in primary care and health care professionals: Development, feasibility, and potential effectiveness studies. *J. Med. Internet Res.* **2023**, *25*, e43293. [CrossRef]
37. Ghatak, S.; Hrithik, P.; Debmitra, G. Voicebot For Mental Disease Prediction and Treatment Recommendation Using Machine Learning. *TechRxiv* **2023**.
38. Davis, S.; Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.* **1980**, *28*, 357–366. [CrossRef]
39. van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. *arXiv* **2016**, arXiv:1609.03499.
40. Mozilla. Common Voice: French Dataset. Available online: https://commonvoice.mozilla.org/fr/datasets (accessed on 6 November 2024).

41. Fadel, W.; Araf, I.; Bouchentouf, T.; Buvet, P.A.; Bourzeix, F.; Bourja, O. Which French speech recognition system for assistant robots? In Proceedings of the 2nd International Conference on Innovative Research in Applied Science, Engineering & Technology (IRASET), Meknes, Morocco, 3–4 March 2022; IEEE Press: New York, NY, USA, 2022; pp. 1–5. [CrossRef]
42. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP, South Brisbane, QLD, Australia, 19–24 April 2015; IEEE Press: New York, NY, USA, 2015; pp. 5206–5210. Available online: https://ieeexplore.ieee.org/abstract/document/7178964/ (accessed on 1 November 2023).
43. Kuchaiev, O.; Li, J.; Nguyen, H.; Hrinchuk, O.; Leary, R.; Ginsburg, B.; Kriman, S.; Beliaev, S.; Lavrukhin, V.; Cook, J.; et al. NeMo: A toolkit for building AI applications using Neural Modules. *arXiv* **2019**, arXiv:1909.09577.
44. NVIDIA; STT_FR_QuartzNet15x5. NVIDIA NeMo Model Catalog. Available online: https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_fr_quartznet15x5 (accessed on 6 November 2024).
45. Majumdar, S.; Balam, J.; Hrinchuk, O.; Balam, J.; Hrinchuk, O.; Lavrukhin, V.; Noroozi, V.; Ginsburg, B. Citrinet: Closing the gap between non-autoregressive and autoregressive end-to-end models for automatic speech recognition. *arXiv* **2021**, arXiv:2104.01721.
46. Huang, Y.; Ye, G.; Li, L.; Gong, Y. Rapid Speaker Adaptation for Conformer Transducer: Attention and Bias Are All You Need. In Proceedings of the Interspeech, Brno, Czechia, 30 August–3 September 2021; pp. 1309–1313.
47. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning ICML'06, Pittsburgh, PA, USA, 25–29 June 2006; ACM Press: New York, NY, USA, 2006; pp. 369–376. [CrossRef]
48. Sharma, R.K.; Joshi, M. An analytical study and review of open source chatbot framework, rasa. *Int. J. Eng. Res.* **2020**, *9*, 1011–1014.
49. Heafield, K. KenLM: Faster and smaller language model queries. In Proceedings of the Sixth Workshop on Statistical Machine Translation, Edinburgh, UK, 30–31 July 2011; pp. 187–197. Available online: https://aclanthology.org/W11-2123.pdf (accessed on 1 November 2023).
50. Chen, S.F.; Goodman, J. An empirical study of smoothing techniques for language modeling. *Comput. Speech Lang.* **1999**, *13*, 359–394. [CrossRef]
51. Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F.M.; Weber, G. Common Voice: A Massively-Multilingual Speech Corpus. *arXiv* **2020**, arXiv:1912.06670.
52. Hirschberg, J.; Manning, C.D. Advances in natural language processing. *Science* **2015**, *349*, 261–266. [CrossRef] [PubMed]
53. Reiter, E.; Dale, R. Building applied natural language generation systems. *Nat. Lang. Eng.* **1997**, *3*, 57–87. [CrossRef]
54. Dhiman, D.B. Artificial Intelligence and Voice Assistant in Media Studies: A Critical Review, SSRN. 2022. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4250795 (accessed on 2 November 2023).
55. Dinesh, R.S.; Surendran, R.; Kathirvelan, D.; Logesh, V. Artificial Intelligence based Vision and Voice Assistant. In Proceedings of the 2022 International Conference on Electronics and Renewable Systems ICEARS, Tuticorin, India, 16–18 March 2022; IEEE Press: New York, NY, USA, 2022; pp. 1478–1483. Available online: https://ieeexplore.ieee.org/abstract/document/9751819/ (accessed on 2 November 2023).
56. Gupta, J.N.; Forgionne, G.A.; Mora, M. *Intelligent Decision-Making Support Systems: Foundations, Applications and Challenges*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2007.
57. Kadali, B.; Prasad, N.; Kudav, P.; Deshpande, M. Home Automation Using Chatbot and Voice Assistant, in ITM Web of Conferences, EDP Sciences, 2020, 01002. Available online: https://www.itm-conferences.org/articles/itmconf/abs/2020/02/itmconf_icacc2020_01002/itmconf_icacc2020_01002.html (accessed on 2 November 2023).
58. Patel, D.; Msosa, Y.J.; Wang, T.; Mustafa, O.G.; Gee, S.; Williams, J.; Roberts, A.; Dobson, R.J.; Gaughran, F. 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.* **2022**, *22*, 100. [CrossRef] [PubMed]
59. Chen, H.; Liu, X.; Yin, D.; Tang, J. A Survey on Dialogue Systems: Recent Advances and New Frontiers. *SIGKDD Explor. Newsl.* **2017**, *19*, 25–35. [CrossRef]
60. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need, Advances in Neural Information Processing Systems. 2017, Volume 30. Available online: https://proceedings.neurips.cc/paper/7181-attention-is-all-you-need (accessed on 2 November 2023).
61. Serban, I.; Sordoni, A.; Bengio, Y.; Courville, A.; Pineau, J. Building end-to-end dialogue systems using generative hierarchical neural network models. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; Available online: https://ojs.aaai.org/index.php/AAAI/article/view/9883 (accessed on 2 November 2023).
62. Cassell, J.; Bickmore, T.; Billinghurst, M.; Campbell, L.; Chang, K.; Vilhjálmsson, H.; Yan, H. 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, PA, USA, 15–20 May 1999; ACM Press: New York, NY, USA, 1999; pp. 520–527. [CrossRef]
63. Følstad, A.; Taylor, C. Investigating the user experience of customer service chatbot interaction: A framework for qualitative analysis of chatbot dialogues. *Qual. User Exp.* **2021**, *6*, 6. [CrossRef]
64. Delorme, J.; Charvet, V.; Wartelle, M.; Lion, F.; Thuillier, B.; Mercier, S.; Soria, J.-C.; Azoulay, M.; Besse, B.; Massard, C.; et al. Natural Language Processing for Patient Selection in Phase I or II Oncology Clinical Trials. *JCO Clin. Cancer Inform.* **2021**, *5*, 709–718. [CrossRef] [PubMed]

65. AI, E. spaCy French Language Models. Available online: https://spacy.io/models/fr (accessed on 6 November 2024).
66. Vincent, M.; Douillet, M.; Lerner, I.; Neuraz, A.; Burgun, A.; Garcelon, N. Using deep learning to improve phenotyping from clinical reports. *Stud. Health Technol. Inform.* **2022**, *290*, 282–286. [PubMed]
67. Honnibal, M.; Montani, I. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. Neural Machine Translation. In Proceedings of the Association for Computational Linguistics, ACL, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 688–697.
68. Bird, S.; Klein, E.; Loper, E. Natural Language Processing with Python, O'Reilly Media. 2009. Available online: https://www.oreilly.com/library/view/natural-language-processing/9780596803346/ (accessed on 9 November 2023).
69. Bocklisch, T.; Faulkner, J.; Pawlowski, N.; Nichol, A. Rasa: Open Source Language Understanding and Dialogue Management. *arXiv* **2017**, arXiv:1712.05181.
70. Gaur, G.; Moh, M.; Zhang, L.; Lin, H. The effects of automatic speech recognition quality on human transcription latency. In Proceedings of the 2016 Conference of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016.
71. Morris, A.C.; Maier, V.; Green, P. From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. In Proceedings of the Eighth International Conference on Spoken Language Processing, Jeju Island, Republic of Korea, 4–8 October 2004.
72. Grinberg, M. *Flask Web Development: Developing Web Applications with Python*; O'Reilly Media Inc.: Sebastopol, CA, USA, 2018.
73. Guazzaroni, G. *Virtual and Augmented Reality in Mental Health Treatment*; IGI Global: Hershey, PA, USA, 2018.
74. Wrzesien, M.; Burkhardt, J.M.; Raya, M.A.; Botella, C. Mixing psychology and HCI in evaluation of augmented reality mental health technology. In Proceedings of the CHI'11 Extended Abstracts on Human Factors in Computing Systems Vancouver, Vancouver, BC, Canada, 7–12 May 2011; ACM Press: New York, NY, USA, 2011; pp. 2119–2124.
75. Le Glaz, A.; Haralambous, Y.; Kim-Dufor, D.H.; Lenca, P.; Billot, R.; Ryan, T.C.; Marsh, J.; DeVylder, J.; Walter, M.; Berrouiguet, S.; et al. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. *J. Med. Internet Res.* **2021**, *23*, e15708. [CrossRef] [PubMed]
76. Niculescu, A.; van Dijk, B.; Nijholt, A.; Li, H.; See, S.L. Making social robots more attractive: The effects of voice pitch, humor and empathy. *Int. J. Soc. Robot.* **2013**, *5*, 171–191. [CrossRef]
77. Funk, B.; Sadeh-Sharvit, S.; Fitzsimmons-Craft, E.E.; Trockel, M.T.; E Monterubio, G.; Goel, N.J.; Balantekin, K.N.; Eichen, D.M.; E Flatt, R.; Firebaugh, M.-L.; et al. A Framework for Applying Natural Language Processing in Digital Health Interventions. *J. Med. Internet Res.* **2020**, *22*, e13855. [CrossRef]
78. Abd-Alrazaq, A.; AlSaad, R.; Aziz, S.; Ahmed, A.; Denecke, K.; Househ, M.; Farooq, F.; Sheikh, J. Wearable artificial intelligence for anxiety and depression: Scoping review. *J. Med. Internet Res.* **2023**, *25*, e42672. [CrossRef]
79. Wadle, L.M.; Ebner-Priemer, U.W.; Foo, J.C.; Yamamoto, Y.; Streit, F.; Witt, S.H.; Frank, J.; Zillich, L.; Limberger, M.F.; Ablimit, A.; et al. Speech Features as Predictors of Momentary Depression Severity in Patients With Depressive Disorder Undergoing Sleep Deprivation Therapy: Ambulatory Assessment Pilot Study. *JMIR Ment. Health* **2024**, *11*, e49222. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.