Hindawi
Journal of Electrical and Computer Engineering
Volume 2023, Article ID 5532967, 12 pages
https://doi.org/10.1155/2023/5532967

Research Article

A Voice-Based Personal Assistant for Mental Health in Kreol Morisien

B. Gobin-Rahimbux,[1] **N. Gooda Sahib**,[1] **N. Peerthy**,[1] **and A. Taylor**[2]

[1]*Software and Information Systems Department, University of Mauritius, Moka, Mauritius*
[2]*Department of Computing and Information Technology, Malawi University of Business and Applied Sciences, Blantyre, Malawi*

Correspondence should be addressed to A. Taylor; ataylor@mubas.ac.mw

Received 19 June 2023; Revised 9 September 2023; Accepted 12 December 2023; Published 28 December 2023

Academic Editor: Islam Abdellah

Copyright © 2023 B. Gobin-Rahimbux et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Voice-based smart personal assistants (VSPAs) are applications that recognize speech-based input and perform a task. In many domains, VSPA can play an important role as it mimics an interaction with another human. For low-resource languages, developing a VSPA can be challenging due to the lack of available audio datasets. In this work, a VSPA in Kreol Morisien (KM), the native language of Mauritius, is proposed to support users with mental health issues. Seven conversational flows were considered, and two speech recognition models were developed using CMUSphinx and DeepSpeech, respectively. A comparative user evaluation was conducted with 17 participants who were requested to speak 151 sentences of varying lengths in KM. It was observed that DeepSpeech was more accurate with a word error rate (WER) of 18% compared to CMUSphinx at 24%, that is, DeepSpeech fully recognized 76 sentences compared to CMUSphinx where only 57 sentences were fully recognized. However, DeepSpeech could not fully recognize any 7-word sentences, and thus, it was concluded that the contributions of DeepSpeech to automatic speech recognition in KM should be further explored. Nevertheless, this research is a stepping stone towards developing more VSPA to support various activities among the Mauritian population.

1. Introduction

Voice-based smart personal assistants (VSPA) are freely available on mobile devices [1–3]. Examples are Siri, Google Now, Cortana, and Alexa among others. These systems enhance the interaction by using speech-based commands. Their aim is to make the interactions with the user more human-like, and the latest advances in AI-enabled speech recognition have contributed a lot to its popularity. Besides mobile devices, voice-activated assistants are found in smart homes, cars, and service encounters and provide support to the elderly and the disabled [4]. It is an undeniable fact that anxiety is one of the most common mental disorders. The number of patients with anxiety is on the rise throughout the world and situations like the COVID-19 pandemic leading to house confinement, and many socioeconomic constraints have produced more cases of anxiety issues. It has been observed that only 2.38 psychiatrists are available for them per 100,000 persons who suffer from psychiatric illnesses [5].

The use of AI-based applications can help to bring a solution to the handling of mild cases, and one of them is voice-based personal assistants (VSPAs) [3, 6]. VSPAs are special applications developed to do a specific task. They take in voice as input and then perform a task as instructed by the voice input. Such systems can provide support and help people by providing 24-hour assistance while medical professionals in the field can attend to more patients with complex issues.

There is a need for research in the field of speech recognition for under-resourced languages to develop applications based on these languages. Kreol Morisien (KM) is a language spoken by the inhabitants of Mauritius, a small island located in the Indian Ocean. Visiting a psychologist is not a common practice on the island. However, the inhabitants also feel stress and anxiety especially due to

COVID-19. A total of 11,038 first-time mental health visits were reported in 2019, of which 6908 were male and 4130 were female [7]. Taking those numbers into account, we can infer that there is a regular rise in mental health disorders in Mauritius. Therefore, to assist people in battling this silent murderer, two models of Kreol voice-based smart assistants for anxiety have been implemented alongside a mobile application. A VSPA will definitely help in bridging the gap and encourage Mauritians to take care of their mental health. This paper presents the work carried out to provide Mauritians with the facility to talk with a voice assistant 24/7 to combat mental health issues using their mobile phone.

VSPAs are usually developed in high-resource languages [8]. For example, the development of the VSPAs in English can be done quite easily due to the availability of large datasets. However, for low-resource languages such as the KM, due to the lack of available audio datasets, it needs to be built from scratch. In this paper, we present the work that has been completed with regard to the development of the VSPA for mental health in Mauritius. The novelty of this work includes the use of DeepSpeech for ASR in a low-resource language. Currently, there have been very few experimentations of speech recognition for low-resource languages using DeepSpeech. Therefore, this work highlights the potential of DeepSpeech for low-resource languages and identifies the challenges that need to be addressed.

Section 2 of the paper gives an overview of VSPA, especially in the context of mental health. Section 3 covers the work done to develop the ASR module in CMUSphinx. Section 4 describes the implementation of the VSPA as well as the user evaluation done. In Section 5, the development of the automatic speech recognition model in DeepSpeech is presented, and Section 7 presents the comparative analysis done between CMUSphinx and DeepSpeech.

2. Literature Review

2.1. Voice-Based Smart Personal Assistants (VSPAs). SPAs have also been integrated into a wide range of consumer markets [9], such as the insurance sector, travel and hospitality industry, finance sector, and smart-home appliances. The market for SPAs has been growing fast due to the integration of artificial intelligence [10]. Winkler et al. [11] define SPAs as sophisticated intelligent programs that respond to users' input by answering questions in natural language, performing actions, or making recommendations. Well-known voice-based SPAs are Replika [12], Amazon's Alexa [13], and Google Siri [8]. By leveraging natural language algorithms, fast Internet, and cloud storage, voice-based SPAs gather users' preferences and customize tasks such as playing favorite music, ordering coffee, booking a taxi, and booking appointments [13]. Chatbots such as Replika act as a personal confidante to the user, by relying on a wide range of data collected about the user through daily interactions and by generating conversations using neural networks and scripted dialogues or conversational flows.

The voice-based SPA technology can be further leveraged by using web services such as Tasker (https://tasker.joaoapps.com/) and IFTTT (https://ifttt.com/) (If This Then That) that provide users with the ability to create their own experiences and integrate the voice-based SPA's capabilities and skills in their mobile applications or smart devices. Using voice-based SPA, the user can personalize their own virtual assistant to allow them to automate social media posts, switch their smart devices off and on, ask them to read the news, and handle events on the calendar [14]. SPAs can be compared using the four-degree measurements such as personification, technological attributes and challenges, and sociability [15]. The degree of personification is the degree to which a customer has configured their own SPA [16]. Technical characteristics and problems are the areas in which the voice-based SPA can perform the tasks provided, such as providing the user with intelligent and smart responses.

SPA skills, also known as virtual assistant capabilities, range from simple skills such as alerts, clocks, or even jokes, to more specialized skills such as playing a particular song, handling calendar activities, and house automation. VSPAs use intent recognition to implement the required skills. Intent recognition is a subfield of artificial intelligence and is a form of the processing of natural language. In addition, intent recognition considers natural language processing, which in any language that has been created spontaneously and not artificially, such as computer code. Goal identification, which is also known as intent classification, is a user's classification and retrieval of intent so that the necessary or applicable capabilities can be implemented by the device. To classify it into an intervention based on what the individual wishes to do, it takes text transcripts or voiced information and uses classification [17]. When the user speaks a sentence, with the help of a speech-to-text engine, the phrase is recognized and then the intent, entity, and facets are identified from the phrase. When the intent has been recognized, the keywords are then filtered and the speech assistant performs the task to be done [18].

Applications such as Alexa and Cortana propose a series of skills that can help in the case of mental health issues. Alexa proposes 8 skills to manage anxiety and reduce stress 7. Yang et al. [19] have compared the responses to 14 frequently asked questions set to voice assistants with postpartum depression. According to the authors, the voice assistants performed well in terms of recognition but none of them managed to go beyond the threshold when it came to providing accurate information on postpartum depression.

Conversational agents have also been integrated into healthcare for mental health [20–22], to prevent suicidal behavior [23], and for age-related depression [24]. However, most of these technologies have been developed for well-resource languages such as English, and adapting them to new languages under resources such as Kreol Morisien comes with challenges and typically requires language work such as developing audio datasets and training new speech recognition algorithms. Automatic speech recognition is a technology that uses vocal waveforms to derive transcription of utterances [25–27]. An ASR system comprises

four main stages, namely, speech analysis, extraction of features, modelling of the ASR, and testing of the system [12]. Some of the tools available to develop ASR models are CMUSphinx (https://cmusphinx.github.io/), KALDI (https://kaldi-asr.org/), and VOXForge (https://www.voxforge.org/). CMUSphinx has been used particularly for training speech models that rely on small voice datasets. Early stages of automatic speech recognition work similar to our work were done for the Dari language in Afghanistan [28] and Arabic languages [29]. Similar to our work, these studies also use the CMU Sphinx toolkit and DeepSpeech to implement an automatic speech voice recognition tool (ASR) for their languages.

2.2. VSPAs for Mental Health. It is known that only a small proportion of individuals struggling with anxiety and mood disorders look forward to receiving professional diagnosis and care [30, 31], with the younger group being especially reluctant to seek help [32]. According to Rickwood et al. [32], such reluctance is often due to them favoring informal support, not having enough faith or experience with mental healthcare, not wanting to be stigmatized, and attempting to tackle the problem on their own. In addition to those individuals, wanting to receive help must overcome numerous other hurdles including the financial cost of therapy, social labelling, and geographical distance to mental healthcare [30].

In recent years, smartphone applications have been perceived to play a crucial role in promoting the delivery and access to mental healthcare and could work as a powerful tool to address the above concerns. Apps can be highly flexible for mental healthcare delivery while, at the same time, being attractive to users. Owning a smartphone is no longer controlled by socioeconomic status or geographical position, being the most favored mode of interaction, especially among youngsters [33]. Smartphones have far-reaching universal networks that can be operated from roughly anywhere in real time, thus enabling users to gain access to digital psychotherapy and support whenever they require. Furthermore, timely detection of mental health issues is essential for prevention and positive health results, which in turn often depends on constant regulation and speedy response [34]. Mobile phones being with the user at all times can help with self-monitoring, giving a unique chance to track behavior and emotional states, in real time and in low-profile way [33, 34]. Moreover, due to their almost immediate response, MH apps can help in recognizing triggers and behavior patterns and, in turn, provide appropriate information based on these real-time variations in effect [34]. Additionally, mobile phone therapy apps, having demonstrated prospects in habit formation [35], can promote user engagement in digital therapy.

It has been observed that there is limited work that discusses the development of voice-based smart personal assistants for healthcare. All applications that have been identified are text-based chatbots developed for smartphones. In a review of mobile chatbot applications for mental health carried out by [36], eleven applications were identified out of which only one had a voice option. An English version was available for all mobile applications. The other languages that were considered were French (2 applications), German (4 applications), Portuguese (2 applications), Romanian (1 application), Spanish (2 applications), Italian (1 application), Russian (1 application), and Arabic (1 application).

The main limitations that were identified based on the literature review were as follows:

(i) The conversational flows for mental health were text-based. They were normally chatbots that would be used to interact with a person as a replacement, and these chatbots would normally provide advice, monitor the progress of the person using the application, and refer to a therapist when needed.

(ii) The interactions with voice-based assistants were mostly command-based and could, therefore, not be considered the same as a conversation with a text-based chatbot for mental health.

(iii) The solutions developed were limited to languages which were very popular. Not much work could be identified for under-resourced languages in this domain.

3. Methodology Used to Develop the VSPA

In this section, the methodology used to develop the VSPA for mental health is described.

3.1. Step 1: Determining the Conversational Flow. To create a conversational flow, the help of a mental health specialist was sought which was then vetted by a second one. Nine conversational flows in Kreol language associated with a specific skill were designed. Every conversational flow (CF) was created with the goal of incorporating the skill of speech recognition and the skill of providing the user with the appropriate feedback audio file. It was decided that voiced-based options would be given to the user whenever the application would ask him a question so that the person would say the options only. In case the person would say something that was not in the option list, then the system would inform the user that it had not understood and would repeat the options to him and request him to choose among the options. This is normally how the conversational flows in text-based chatbots are set. Figure 1 shows the various conversation flows. The inputs and outputs of each conversational flow are presented in Table 1.

The emphasis of the study was to develop the first shell for the voice recognition module in Kreol Morisien for mental health, and the focus was on the rate of recognition of the conversation so that the right skills are triggered. The voice recognition module is expected to act as a foundation for an enhanced version that would consider other aspects as emotions and also open speech where the user will be allowed to say things by himself instead of being requested to

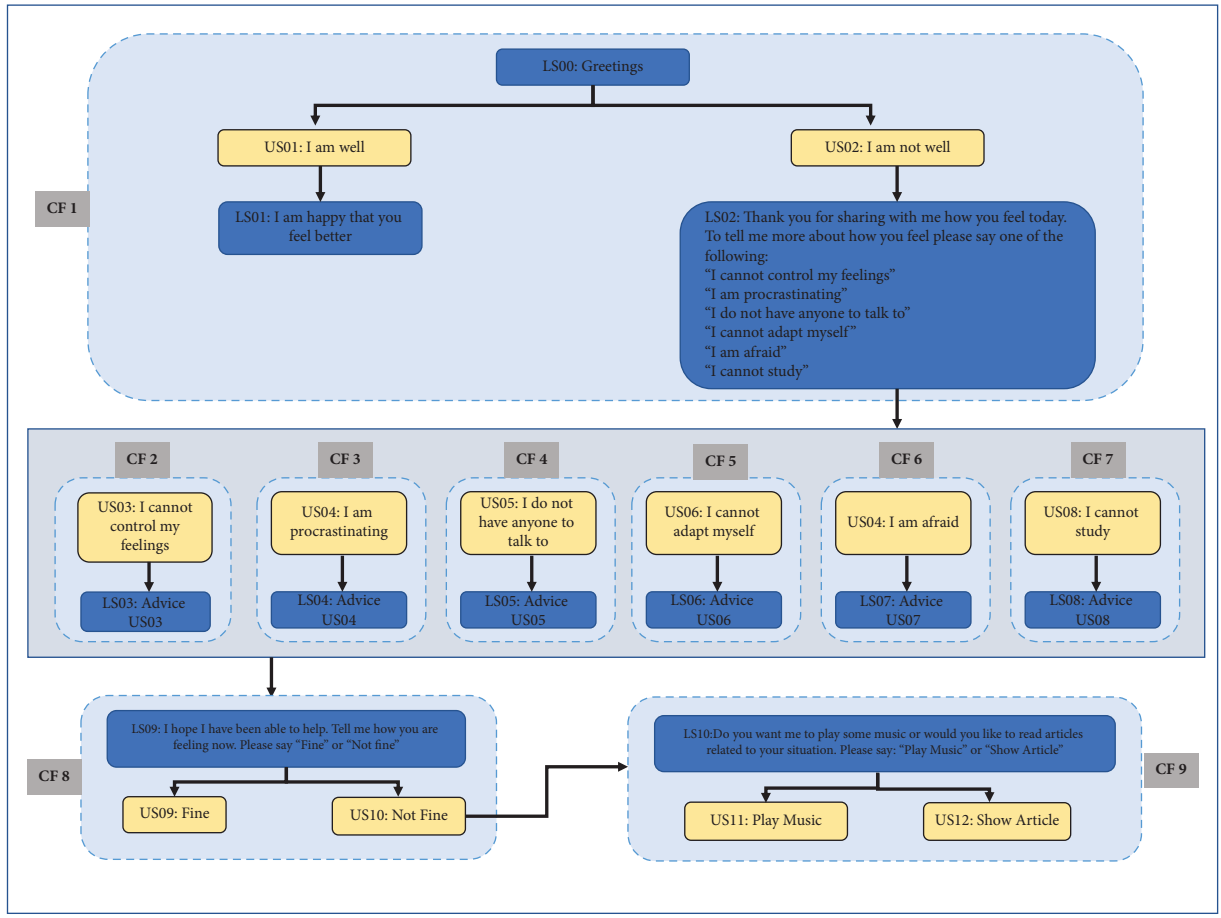

Figure 1: Conversational flow built in Kreol Morisien.

choose between the range of options. Also, a set of conversational flows can be used to develop similar applications for mental for other low-resource languages.

3.2. Step 2: Building the Audio Dataset. A script consisting of a list of sentences was given to the participants to be read aloud and then recorded. For symptoms and behaviors specifically related to anxiety and stress, a series of psychology and psychiatry journal papers, existing clinical scales (such as GAD-7, Beck Anxiety Inventory, Hamilton Anxiety scale, Perceived Stress Scale, and Depression Anxiety Stress Scale), and the Diagnostic and Statistical Manual of Mental Disorders-5 (DSM 5) were reviewed. To ensure uniformity across the study materials and information gathered, they were translated into Kreol Morisien using the Diksioner Morisien [37] and in consultation with a counsellor.

To make the system inclusive, close attention was paid to cover for the speaker's variability by including different variations in word pronunciation. Based upon the design of the virtual assistant, the mental health script was split into 5 sections including greetings and farewell, anxiety and stress symptoms, user responses, system responses, screening questions, and scoring, and breathing exercises. Once completed, a group of psychologists and counsellors was approached to validate the content of the script, of whom four responded favorably. The first version of the script was reviewed by two psychologists, one counsellor, and one academic, and necessary changes were made. The script was validated upon its third revision, and we were then allowed to proceed with the recordings.

Speech-based systems are often highly sensitive to speaker and environment variability. Each speaker differs in age, dialect, personal style, education, gender, and health. Even when interspeaker differences are dismissed, no same speaker can generate the exact utterance twice. Word error rates (WER) are dependent upon speech accent and rate, with higher speaking rates and heavier accents yielding greater WER. Situational variables, for instance, background noise, device noise, position of the microphone during recording, and involuntary sounds such as lip smacks and breathing, may also influence the system's performance.

Taking into consideration all points mentioned above, a total of 12 participants, including 6 males and 6 females, aged 18–40 years, were recruited using a convenience sampling method. To account for accent variability, the number of participants from urban and rural regions was evenly distributed across both genders, with attention given to recruit participants in the rural group from regions mapping the four cardinal points of the island. Participants were given the script and asked to read it aloud, taking

Table 1: Description of conversational flows.

CF*	Skill	Description
CF1	Greeting and responding to greetings	In this conversational flow, the system greets the user and asks the user whether he/she is well. In case the user is well, then the system gives an audio response stating that he is very happy and the person is better. In case the person is not well then, the system thanks the user for sharing his state of mind and requests that the user choose between a number of options
CF2	Advise on how to control feelings	User input is "I cannot control my feelings." The system gives a series of recommendations and activities to do, e.g., accept one's feelings or meet a friend. The user is then invited to say whether the system has been able to help
CF3	Advise on procrastination	User input is "I am procrastinating." The system gives motivational and organizational advice, and then, the user is then invited to say whether the system has been able to help
CF4	Advise on what to do when the person does not have anyone to talk to	User input is "I do not have anyone to talk to." The system advises on how to socialize
CF5	Advise on how a person can adapt to a situation	User input is "I cannot adapt to this current situation." The system gives advise on how to deal with difficult situations/
CF6	Advise on fear	User input is "I am afraid." Systems give instructions for deep breathing
CF7	Advise on studies	User input is "I cannot study." Systems gives advice on good sleep habits and how to study
CF8	Ask if advise was helpful	User input is "fine" or "not fine." System will request the user to say only these two options
CF9	Play music/show article	If the user input was "not fine," then the system will ask whether the user wants it to play music or show an article, and based on the user's selection, the system will do one of the tasks

*CF: conversational flow.

a short pause after each sentence. Audio was recorded using a professional microphone connected to a desktop computer. Most recordings were done in a computer laboratory with a mild to moderate level of noise and then saved in the CMU Sphinx software-compatible format (WAV 16-bit PCM, uncompressed mono-channel).

Participants' recordings were coded in the following format: "speaker gender, speaker number," for example, M1 and F2. Audacity (https://www.audacityteam.org/) was used to clean and segment the long audio files. For recordings deemed too noisy, background noise was removed where possible using audacity's noise reduction option following which the file was split into tracks. Each WAV file was then played and matched with its phrase counterpart from the script. Only complete and well-spoken phrases were kept. The transcription was saved in a text document whereby each phrase was linked to its corresponding WAV file. A total of 3751 audio files were obtained.

3.3. Step 3: Building the ASR Model in CMU Sphinx. CMU Sphinx, released on 5th August 2015, is a popular speech recognition toolkit that includes a number of tools for creating voice applications. A selection of speech recognition systems is included in the CMUSphinx. A front end, linguist, AM, dictionary, and LM make up the CMU's main architecture. The front end consists of the GUI from which the voice is entered into the device. The LM consists of the AM, the dictionary, and the model of language [38]. In addition, the LM consists of a mathematical package in which it is possible to find the likelihood of a term occurring. The dictionary consists of the same word phonetic structure that can be used to distinguish between pronunciations of the same words [39]. The AM consists of a mathematical representation of sound waves at various pitches.

CMU Sphinx was preferred as a tool for development as it allowed to create models from scratch and could be easily integrated in Android. Moreover, a very large dataset is not required to build the model. To create a CMU Sphinx language model, it is crucial to prepare a text file with all transcriptions in the Kreol language. A reference text was prepared to be utilized in the language model generation. The vocabulary file and language models were generated. Using a set of sample speech signals, the trainer learns the parameters for the sound unit model. The database holds the data needed to extract statistics from speech in the form of an acoustic model. Two dictionaries were built, namely, the language phonetic dictionary, and the second was the filler dictionary. The phonetic dictionary of a language is a valid list of words that are mapped to segments of sound units. The filler dictionary is the list of nonspeech sounds. The trainer then consults the dictionary to determine the sound unit sequence that corresponds to each sound signal and its transcription.

The training of the CMU Sphinx model was carried out in three phases. The first phase entails cleaning up the directories to verify that no outdated models are there. The AM's flat initialization is the second phase, in which all mixture weights are set to be equal for all states and all state transition probabilities are set to be equal. The Baum–Welch method, which does forward and backward reestimation, is used in the last phase where it performs several "passes" of the Baum–Welch reestimation over the training data since this is an iterative reestimation procedure. Each of these rounds, or passes, yields a slightly improved set of models for the CI phones. The training was successfully completed at iteration 6.

3.4. Step 4: Development of the Mobile Application. Figure 2 shows the 2-tier system architecture for the VSPA named Liza. The data layer involves the ASR module and the intent recognition process. The ASR contains the AM and LM of the CMU Sphinx SST engine, while the intent recognition involves the decoder and top-scoring intent to filter keywords. Finally, an audio response is then generated by the VSPA which is carried out in the presentation layer. Figure 2 shows the detailed 2-tier system architecture diagram of the mobile application. The user is requested to input his name if he is a first-time user. After the welcome message, the user is asked how he is feeling. The application waits for up to 10 seconds to get an answer. Once the answer is obtained, then the tasks are carried out if the answer is recognized by the system.

In the intent recognition module, arrays are declared for extracting entities, intent, and verbs from the hypothesis. To classify intents and entities, a classifier approach is used. Initializing categorized arrays sets up the classifier. The classifier method calculates the probabilities for an entity, intent, and verb that the user might have said.

3.5. Step 5: Evaluating the Usability of the VSPA. For the user evaluation, 50 participants installed the mobile application on their mobile phones and interacted with the VSPA. Convenience sampling was used as it involved the people who were the most available, and it was a quick and low-cost technique to acquire preliminary data. However, to make sure that all age groups get to use the mobile application, people of different age groups and genders were considered. As the Kreol languages may differ when it is being spoken by people who live in urban and rural regions, participants were chosen from both regions.

To be able to analyze the data received by users more accurately, all fifty participants were asked to do all nine conversational flows. The demographics of the 50 participants are shown in Table 2. The conversational flows (CF) were provided to the participants who were requested to say sentences only related to a conversational flow. The users were then asked to state whether the output given by Liza was as expected for each conversational flow. For example, for conversational flow CF1, the participants were given the scripts to read as input, and then, the output of the VSPA was checked to see whether it was according to the conversation flow. If the output was not as required at any step of the conversation flow, then it would mean that the sentences were not recognized properly. Figure 3 presents the responses obtained for each CF. In the conversational flow CF 1, 35 participants said that the response of Liza was according to the conversational flow while 15 said that the response was not

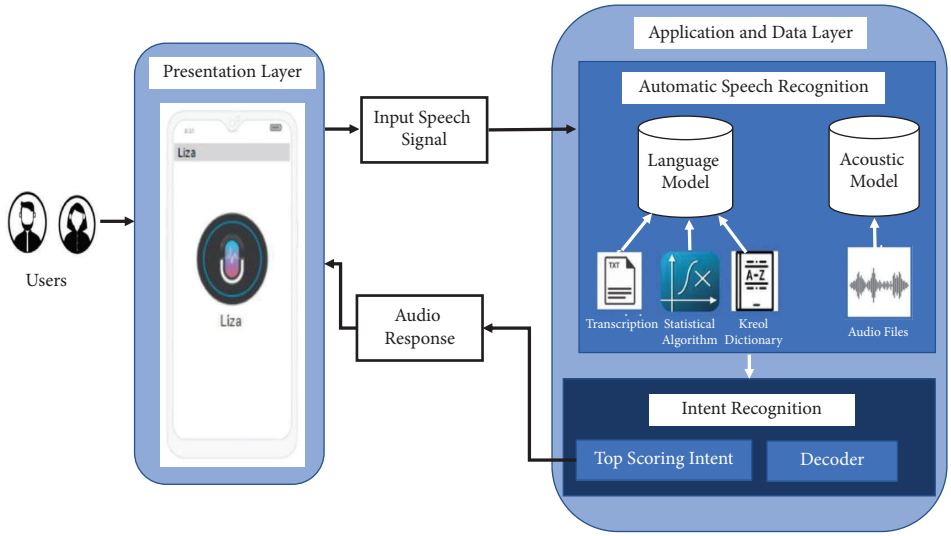

Figure 2: 2-Tier system architecture diagrams of Liza, the voice-activated personal assistant for anxiety.

Table 2: Demographics of the participants.

Demographics	Sample
Age	8% less than 18 52% between 18 and 29 24% between 30 and 40 16% above 40
Gender	40% male and 60% female
Region	38% rural and 62% urban

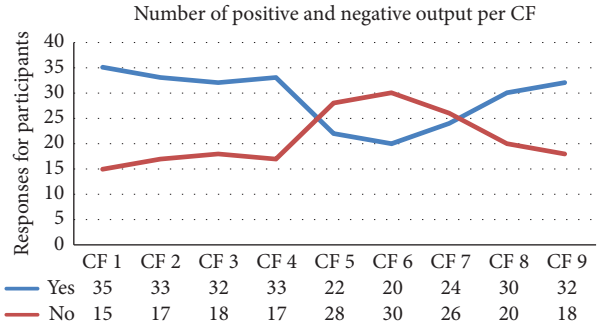

Figure 3: Number of positive and negative outputs for each CF.

which means that Liza did not recognize the input of these 15 participants for CF1. CF1 has the highest number of good output. CF6 has the highest number of wrong output, with only 20 respondents stating that Liza was able to give the proper output for this conversational flow. On average, the overall percentage of good responses is 58%. There is, therefore, definitely a need to improve the recognition rate.

4. Development of the ASR Model Using DeepSpeech

The next step of the work was to develop another ASR model with DeepSpeech so as to compare which model would be more accurate. Launched in May 2016, DeepSpeech is currently regarded as the most efficient tool for speech-to-text recognition. DeepSpeech (https://www.microsoft.com/en-us/research/project/mavis/) uses an end-to-end deep neural network architecture. This section will document how the neural network structure was implemented from scratch in DeepSpeech and how it was finally trained using machine learning techniques.

4.1. Building the Language and Acoustic Model. This step involved the creation of a scorer. A scorer is made up of two parts: a KenLM LM and a true data structure that contain all words that are expected to be used during the interaction. It is very crucial to create the scorer file as it is used alongside the AM to be able to produce speech-to-text interaction. To create a scorer, the text data that correspond to the intended use case were structured in a text file with one sentence per line. A tree, also known as a digital tree, is then created from the file. This involved defining a set of parameters for the LM, building the ARPA model, and generating the LM binary.

The KenLM was created by CMake. Once the build files for KenLM were created, the LM binary and a vocabulary text file were generated. When the tree was built, a vocabulary file consisting of unique and most frequent terms was output in the output directory. The score was then generated using the alphabet text file which contains all characters of the Kreol Morisien language which are found in the vocabulary file.

Once the scorer file has already been generated, the next step is to create the AM. Although the scorer file can be created later, it is recommended that the scorer file be used to generate the AM as well. The audio files were converted to the required format. Then, three CSV files containing transcripts for each audio file with UTF8 encoding were created. Three directories are to be created, namely, train, test, and dev, where train.csv, test.csv, and dev.csv were also created. The content of each of the CSV files must start with three of the following fields:

(i) wav_filename will contain the absolute or relative path where the audio file is located.

(ii) wav_filesize will contain the size of the samples in bytes, which is used to sort the data before training. Integer is anticipated.

(iii) Transcript will contain the transcription for each audio file.

The two options that are available to build the AM in DeepSpeech are fine-tuning and transfer learning. It is crucial to train the AM using the right method, else the model output can either be useless or inaccurate. Fine-tuning is the process of modifying a model that has already been conditioned for one task to make it perform a second equivalent task. For example, an AM that recognizes English can be fine-tuned to recognize an Indian accent for English as well. It splits the functionality of a neural network by making minor changes to the design, details, or learning process. Transfer learning is a strategy in which you freeze certain layers and then train the remaining layers to match your needs and goals. To use transfer learning in DeepSpeech, it is crucial to have a different kind of language. The transfer learning model undergoes various modifications, including the removal, initialization with new values, or merging of layers from a pre-existing model. Additional layers may be introduced, and parameters are reinitialized to accommodate the new target alphanet. Gradient descent is used to update the layers.

Upon closer analysis of both methods available, the transfer learning approach was preferred. The layers were initialized and modified for transfer learning. Parameters to generate the AM were defined, and the AM was optimized and tested. The best word error rate was obtained. Finally, the model was exported so that it could be used.

5. Comparative Analysis between CMU Sphinx and DeepSpeech

In this section, the comparative analysis between the two models previously built is presented.

5.1. Speech Engine Model. The speech engine model converts any standard language model submitted to the knowledge base into a graph. DeepSpeech was parameterized to utilize a bidirectional RNN implemented using TensorFlow, which implies it needs to have all inputs before it can start doing anything helpful. RNNs are neural networks that "remember." They accept as input not only the next word in the recognition, but also a state that grows over time, and utilize this state to capture time-dependent patterns. Table 3 provides an overview of the model size of each speech engine.

5.2. Training of Speech Engine Model. CMU Sphinx is significantly easier to train. This is due to CMU Sphinx reliance on external knowledge sources such as a phonetic lexicon. It can even recognize odd names, such as Kreol names, by simply adding a word to the lexicon and language model. End-to-end systems, that is, the DeepSpeech model, are difficult to train since they demand a large amount of data and computational resources. Many attempts are required to train the DeepSpeech model to maximize training hyperparameters such as learning rate. In addition, it cannot create a new word with an uncommon letter combination because the system will never learn to recognize it because it has not seen it before.

Table 3: Model size of each speech engine.

	CMU Sphinx	DeepSpeech
Size of checkpoints generated	No checkpoint generated	3.17 GB
Size of AM	39.4 KB	180 MB
Size of LM	51.2 KB	23.8 KB

5.3. Evaluation of DeepSpeech and CMU Sphinx SST Engine. The word error rate (WER) is a standard metric for comparing the accuracy of speech recognition transcripts [40]. The WER is calculated using a measurement known as the "Levenshtein distance" and is calculated in terms of the number of substitutions, insertions, and deletions for a number of words that are spoken as shown in the following equation:

$$\text{WER} = \frac{\text{Substitutions} + \text{Insertions} + \text{Deletions}}{\text{Number of words that were actually said}}. \quad (1)$$

The lower the WER, the reeffective the ASR program is at understanding speech. As a result, a higher WER also implies lower ASR accuracy. WER may be influenced by a number of factors that are not inherently related to the capabilities of the ASR technology, namely, (1) the accuracy of the recordings, (2) the consistency of the microphone, (3) the pronunciation of the speaker, (4) noise in the background, (5) names, places, and other proper nouns with unusual spellings, and (6) words that are technical or industry-specific. Figure 4 shows the average WER for healthcare conversational speech for various ASR engines [40]. The value ranges from 35% (Microsoft Mavis (https://www.microsoft.com/en-us/research/project/mavis/)) to 65% for DeepSpeech.

5.4. Methodology Used to Compare DeepSpeech and CMU Sphinx. For the evaluation of DeepSpeech and CMU Sphinx, seventeen participants (Table 4) were chosen to do the speech-to-text interaction for both DeepSpeech and CMU Sphinx. To choose the participants, convenience sampling was chosen because it includes the most accessible population and is a rapid and low-cost method of gathering early data. However, to better assess the accuracy of the voice-based assistant, different participants having different age groups and genders were chosen as the tone of the speaker will differ and the skill of speech recognition will also differ. The accents of people living in urban regions differ from people living in rural regions. There is a slight change

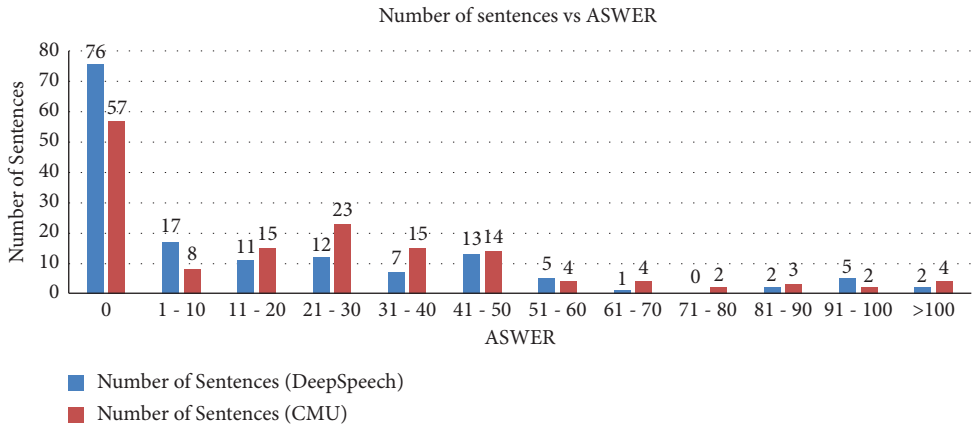

Figure 4: Number of sentences vs. average sentence WER percentage.

Table 4: Details of participants who were part of the comparative evaluation of CMU Sphinx and DeepSpeech.

Age	Gender		Native region	
	Male	Female	Urban	Rural
<18	3	0	3	0
18–29	2	3	3	2
30–40	2	2	2	2
>40	2	3	5	0

in the Kreol language that they speak, but as the model was trained in both accents, this also was taken into consideration and participants from different regions were also selected. To be able to analyze the WER effectively, all seventeen participants were requested to say 151 sentences with the DeepSpeech model as well as with the CMU Sphinx model. The WER for each sentence for each participant was calculated. The average WER of the sentence was then calculated by averaging the values obtained for the 17 participants as discussed in Section 6.

6. Results

The average WER per sentence (ASWER) was calculated by averaging WER obtained for each sentence for all 17 participants as shown in the following equation:

$$\text{ASWER}_x = \left(\frac{1}{n}\sum_{i=0}^{n}\text{SWER}_i\right) \times 100, \tag{2}$$

where n = number of participants

As shown in Figure 5, 76 sentences were fully recognized by the DeepSpeech model compared to CMU Sphinx which recognized only 57 sentences. DeepSpeech had 17 sentences with WER between 1% to 10% and 11 sentences with WER between 11% to 20%. In comparison, CMU Sphinx had 8 sentences with WER between 1% to 10% and 15 sentences with WER between 11% to 20%. From the diagram, it can be seen that CMU has more sentences with WER across different ranges except for 51–60 and 91–100. Table 5 shows the breakdown of the recognition of the sentences. DeepSpeech performed better than CMU for sentences containing 3, 4, and 5 words, respectively.

An analysis of the WER with regard to the length of the sentence was also carried out. Figure 6 shows the distribution of the average sentence WER for DeepSpeech, while Figure 7 shows the distribution for CMU Sphinx. It can be seen from the trendline that the ASWER for DeepSpeech decreased considerably as the length of the sentence increased. The recognition became better compared to CMU Sphinx for the same sentences.

The WER rate of the models was calculated by averaging the ASWER of the 15q sentences as shown in the following equation:

$$\text{WER}_m = \left(\frac{1}{s}\sum_{x}^{1}\text{ASWER}_x\right), \tag{3}$$

where s = number of sentences

The WER for DeepSpeech was 18% while that of CMU Sphinx was 24%. Therefore, the DeepSpeech model was found to be more accurate. It was not possible to compare the results with similar existing work as this is the first attempt to develop a conversational agent in KM for mental health. Figure 7 shows the results obtained for ASRs in healthcare where the range for the WER is between 35% and 65%. The clinical scenarios were not the same, and the language used was English, with native English speakers testing the system. The only other work published in the health domain in KM [41] reports a WER of 17.91% for CMU Sphinx when the evaluation was performed in environments of varying levels of noise. However, the sentences were related to medical symptoms and therefore relatively short.

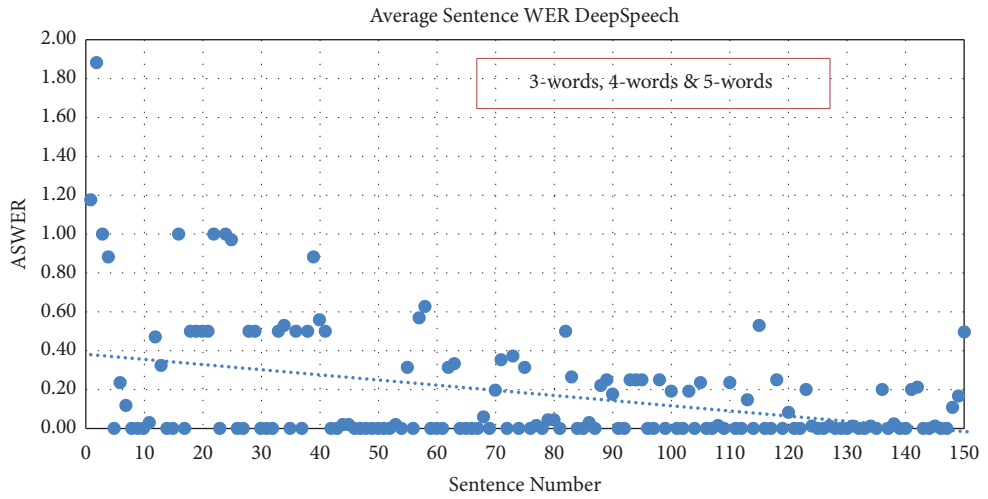

Figure 5: Average sentence WER for DeepSpeech.

Table 5: Recognition of sentences of different lengths.

No. of words in sentence	Number of sentences	DeepSpeech	CMU
1 word	6	1	0
2 words	35	14	16
3 words	34	21	16
4 words	43	22	14
5 words	28	17	9
6 words	3	1	1
7 words	2	0	1
	151	76	57

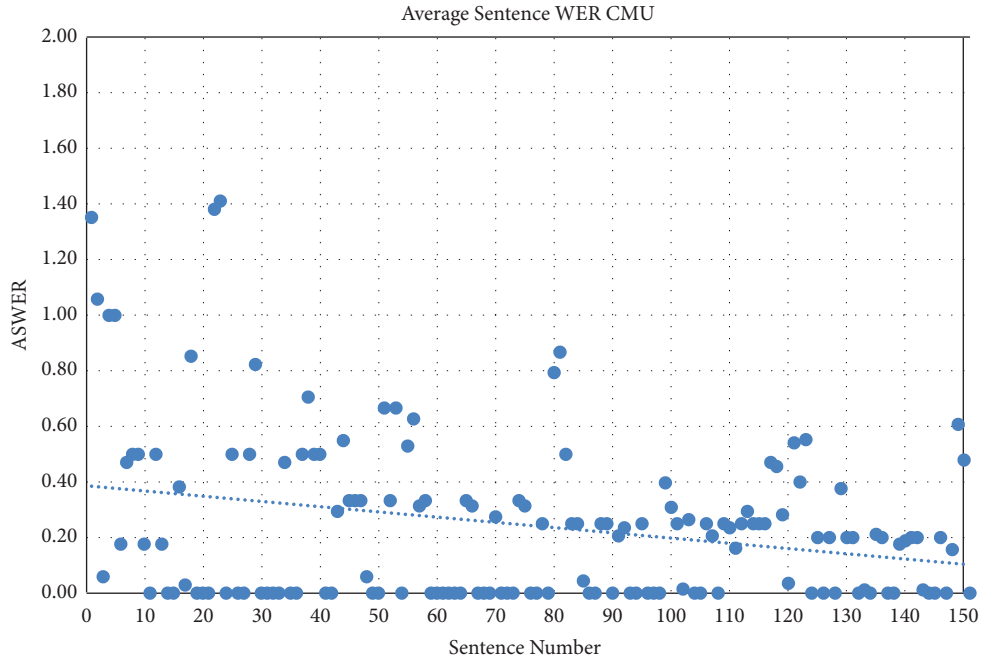

Figure 6: Sentence WER for CMU Sphinx.

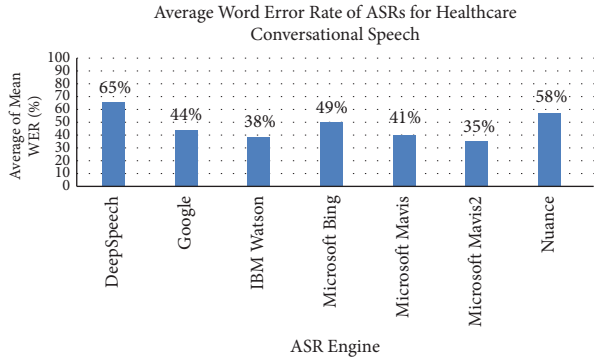

Figure 7: Average WER for ASRs for healthcare conversational speech.

7. Conclusion

This paper discusses the work carried out with regard to the development of the VSPA for mental health in Kreol Morisien. Kreol Morisien is a low-resource language that is spoken by people living in the Republic of Mauritius, and the lack of available resources makes natural language processing in the language challenging. In this work, an audio dataset in KM is created based on a conversation flow between a mental health practitioner and a person having anxiety issues. The safe interaction recommended by the mental health practitioner was taken into consideration, and nine conversational flows were designed which involved four skills that the VSPA would contain. Every conversational flow was created with the goal of incorporating the skill of speech recognition and the skill of providing the user with the appropriate feedback audio file.

A speech recognition module was developed using CMU Sphinx (statistical approach) as well as DeepSpeech (AI approach), and the accuracy of both approaches was compared. Currently, there have been very few experimentations of speech recognition for low-resource languages using DeepSpeech. Therefore, this work highlights the potential of DeepSpeech for low-resource languages and identifies the challenges that need to be addressed.

An additional contribution of this research is the evaluation of the two speech recognition modules whereby 17 participants were asked to use the system with predefined conversations. In this respect, the word error rate (WER) for DeepSpeech was found to be the lowest. This work also experimented with DeepSpeech with relatively small datasets. However, the results are encouraging and lead us to believe that with higher amounts of audio data, DeepSpeech can significantly impact on the development of voice-based systems for low-resource languages.

In the future, it is expected that larger datasets will be created to be experimented with DeepSpeech. Additionally, this work looks at Kreol Morisien, the native language of Mauritius. However, in many parts of the world such as Africa, initiatives are ongoing to create accessible technology for their populations. An important aspect of accessibility remains for voice-based interaction, and therefore, DeepSpeech can be experimented with other such low-resource languages. The work also focused on the accuracy of the recognition. Other aspects like tone recognition to enhance the interaction between the user and the VSPA will be considered.

Data Availability

The audio dataset used to support the build of the automatic speech recognition module in this study is available from the main author upon request Email: b.gobin@uom.ac.mu.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This research was funded by the University of Mauritius (RFS-A)–Project no. RA020.

References

[1] J. S. Edu, J. M. Such, and G. Suarez-Tangil, "Smart home personal assistants: a security and privacy review," *ACM Computing Surveys*, vol. 53, no. 6, pp. 1–36, 2020.

[2] M. Dubiel, M. Halvey, and L. Azzopardi, "A survey investigating usage of virtual personal assistants," 2018, https://arxiv.org/abs/1807.04606.

[3] A. Coskun-Setirek and S. Mardikyan, "Understanding the adoption of voice activated personal assistants," *International Journal of E-Services and Mobile Applications*, vol. 9, no. 3, pp. 1–21, 2017.

[4] A. Pradhan, K. Mehta, and L. Findlater, "Accessibility came by accident' use of voice-controlled intelligent personal assistants by people with disabilities," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, Montreal, QC, USA, April 2018.

[5] World Health Organisation, "Mental health atlas 2017," 2018, https://www.who.int/publications-detail-redirect/9789241514019.

[6] R. Winkler, C. Büchi, and M. Söllner, *Improving Problem-Solving Skills with Smart Personal Assistants: Insights from a Quasi Field experiment*, ICIS, New York, NY, USA, 2019.

[7] Health Statistics Unit, "Ministry of health and wellness," 2020, https://main.mohfw.gov.in/.

[8] M. B. Hoy, "Alexa, Siri, Cortana, and more: an introduction to voice assistants," *Medical Reference Services Quarterly*, vol. 37, no. 1, pp. 81–88, 2018.

[9] R. Knote, A. Janson, L. Eigenbrod, and M. Söllner, "The what and how of smart personal assistants: principles and application domains for is research," *Multikonferenz Wirtschaftsinformatik (MKWI)*, vol. 25, 2018.

[10] mordorintelligence, "Intelligent virtual assistant (IVA) market analysis- industry report- trends, size and share," 2023, https://www.mordorintelligence.com/industry-reports/intelligent-virtual-assistant-market.

[11] R. Winkler, M. Söllner, and J. M. Leimeister, "Enhancing problem-solving skills with smart personal assistant technology," *Computers & Education*, vol. 165, Article ID 104148, 2021.

[12] T. Gulzar, A. Singh, D. K. Rajoriya, and N. Farooq, "A systematic analysis of automatic speech recognition: an

overview," *International Journal of Current Engineering and Technology*, vol. 4, no. 3, pp. 1664–1675, 2014.

[13] R. Knote, M. Söllner, and J. M. Leimeister, "Towards a pattern language for smart personal assistants," in *Proceedings of the Conference on Pattern Languages of Programs (PLoP)*, Coimbatore, India, June 2018.

[14] A. Srinivasan and A. N. Madheswari, "The role of smart personal assistant for improving personal healthcare," *International Journal of Advanced Engineering, Management and Science*, vol. 4, no. 11, pp. 769–772, 2018.

[15] T. Gulzar, A. Singh, and S. Vijay, "An improved endpoint detection algorithm using bit wise approach for isolated, spoken paired and Hindi hybrid paired words," *International journal of computer applications*, vol. 92, no. 15, pp. 1–12, 2014.

[16] N. Mallat, V. Tuunainen, and K. Wittkowski, "Voice activated personal assistants–consumer use contexts and usage 16 behavior," in *Proceedings of the Technology Research, Education, and Opinion (TREO)*, Americas Conference on Information Systems (AMCIS'17), Boston, MA, USA, August 2017.

[17] C. Marshall, "What is intent recognition and how can I use it?" 2020, https://medium.com/mysuperai/what-is-intent-recognition-and-how-can-i-use-it-9ceb35055c4f.

[18] D. Mohan, "Joint intent classification and entity recognition for conversational commerce," 2019, https://medium.com/walmartglobaltech/joint-intent-classification-and-entity-recognition-for-conversational-commerce-35bf69195176.

[19] S. Yang, J. Lee, E. Sezgin, J. Bridge, and S. Lin, "Clinical advice by voice assistants on postpartum depression: cross-sectional investigation using apple Siri, Amazon Alexa, Google assistant, and microsoft Cortana," *JMIR mHealth and uHealth*, vol. 9, no. 1, Article ID e24045, 2021.

[20] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "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 mental health*, vol. 4, no. 2, p. e19, 2017.

[21] A. Miner, "Conversational agents and mental health: theory-informed assessment of language and affect," in *Proceedings of the Fourth International Conference on Human Agent Interaction*, pp. 123–130, Singapore, July 2016.

[22] L. Ring, T. Bickmore, and P. Pedrelli, "An affectively aware virtual therapist for depression counseling," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI) Workshop on Computing and Mental Health*, New Orleans, LA, USA, May 2016.

[23] J. Martínez-Miranda, A. Martínez, R. Ramos et al., "Assessment of users' acceptability of a mobile-based embodied conversational agent for the prevention and detection of suicidal behaviour," *Journal of Medical Systems*, vol. 43, no. 8, p. 246, 2019.

[24] J. Striegl, M. Gotthardt, C. Loitsch, and G. Weber, "Investigating the usability of voice assistant-based CBT for age-related depression," in *Proceedings of the Computers Helping People with Special Needs: 18th International Conference, ICCHP-AAATE 2022*, Lecco, Italy, July 2022.

[25] M. Zajechowski, "Automatic speech recognition (ASR) software- an introduction," 2014, https://usabilitygeek.com/automatic-speech-recognition-asr-software-an-introduction/.

[26] P. C. Raut and S. U. Deoghare, "Automatic speech recognition and its applications," *International Research Journal of Engineering and Technology*, vol. 3, no. 5, pp. 2368–2371, 2016.

[27] N. Najkar, F. Razzazi, and H. Sameti, "A novel approach to HMM-based speech recognition systems using particle swarm optimization," *Mathematical and Computer Modelling*, vol. 52, no. 11–12, pp. 1910–1920, 2010.

[28] D. Yousufzai, "Early stages of automatic speech recognition (ASR) in non-English speaking countries and factors that affect the recognition process," *American Journal of Neural Networks and Applications*, vol. 7, no. 1, pp. 15–22, 2021.

[29] H. Satori, H. Hiyassat, M. Haiti, and N. Chenfour, "Investigation Arabic speech recognition using CMU sphinx system," *The International Arab Journal of Information Technology*, vol. 6, no. 2, 2009.

[30] D. Bakker, N. Kazantzis, D. Rickwood, and N. Rickard, "Mental health smartphone apps: review and evidence-based recommendations for future developments," *JMIR mental health*, vol. 3, no. 1, p. e7, 2016.

[31] R. Mojtabai, M. Olfson, N. A. Sampson et al., "Barriers to mental health treatment: results from the national comorbidity survey replication," *Psychological Medicine*, vol. 41, no. 8, pp. 1751–1761, 2011.

[32] D. Rickwood, F. P. Deane, C. J. Wilson, and J. Ciarrochi, "Young people's help-seeking for mental health problems," *Australian e-journal for the Advancement of Mental health*, vol. 4, no. 3, pp. 218–251, 2005.

[33] V. Harrison, J. Proudfoot, P. P. Wee, G. Parker, D. H. Pavlovic, and V. Manicavasagar, "Mobile mental health: review of the emerging field and proof of concept study," *Journal of Mental Health*, vol. 20, no. 6, pp. 509–524, 2011.

[34] N. Rickard, H.-A. Arjmand, D. Bakker, and E. Seabrook, "Development of a mobile phone app to support self-monitoring of emotional well-being: a mental health digital innovation," *JMIR mental health*, vol. 3, no. 4, p. e49, 2016.

[35] A. Oulasvirta, T. Rattenbury, L. Ma, and E. Raita, "Habits make smartphone use more pervasive," *Personal and Ubiquitous Computing*, vol. 16, no. 1, pp. 105–114, 2012.

[36] A. Ahmed, N. Ali, S. Aziz et al., "A review of mobile chatbot apps for anxiety and depression and their self-care features," *Computer Methods and Programs in Biomedicine Update*, vol. 1, Article ID 100012, 2021.

[37] A. Carpooran, *Diksioner Morisien: Premie Diksioner Kreol Monoleng, 2em Edision*, Koleksion Text Kreol Sainte Croix, Maurice: Les Éditions Le Printemps, Mauritius, East Africa, 2011.

[38] A. Ankit, "Acoustic speech recognition for Marathi language using sphinx," *ICTACT Journal on Communication Technology*, vol. 7, no. 3, pp. 1361–1365, 2016.

[39] A. G. Adami, "Automatic speech recognition: from the beginning to the Portuguese language," in *Proceedings of the 9th International Conference on Computacional Processing of the Portuguese Language*, New Jersey, NY, USA, August 2010.

[40] M. Gervitz, "What is word error rate (WER)?- deepgram blog," 2018, https://blog.deepgram.com/what-is-word-error-rate/.

[41] N. Gooda Sahib-Kaudeer, B. Gobin-Rahimbux, B. S. Bahsu, and M. F. A. Maghoo, "Automatic speech recognition for kreol morisien: a case study for the health domain," in *Proceedings of the International Conference on Speech and Computer*, Springer, Berlin, Germany, June 2019.