LARRY GOYEAU

Technical dossier for innovation tax credit approval (crédit impôt innovation)

Year 2024

SIREN : 898 663 158

NACE : 6201Z

Address : 475 Chemin des communaux 38190 Bernin

Creation date: 01/16/2024

Summary

1 Presentation of LARRY GOYEAU

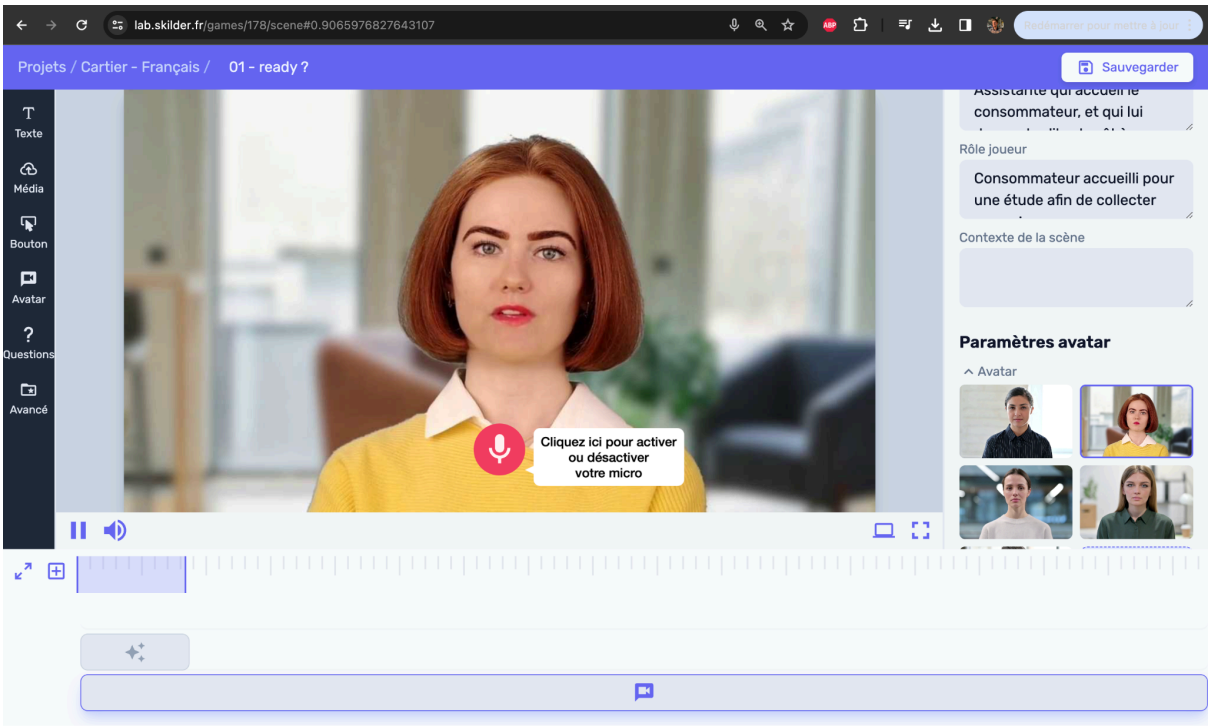
Larry GOYEAU, an engineer and founder of the company bearing his name, specializes in developing artificial intelligence solutions applied to image and sound processing. He is known for his expertise in creating web applications that integrate AI models for real-time processing. His work also includes developing interactive user interfaces aimed at enhancing user experience. In his previous experiences, Larry GOYEAU had the opportunity to develop a body movement analysis tool for medical purposes. The tool works by using a camera to film the user doing exercises and is capable of estimating their movement in 3D. He has also worked on the automatic detection of diseases from X-rays, which has allowed him to deepen his skills in image processing and data science.

2 Project

Name of the Project	**Avatar intervieweur**		
Start date	**September 18, 2023**	**End date**	**End of 2024**
Launched in September 2023 at sweeeft.ai, the project aims to develop an avatar robot programmed to conduct interviews with candidates. The avatar, designed to automate recruitment processes, can dynamically adjust to respondents' answers while following predefined guidelines. Its design allows for extensive customization, offering recruiters the ability to modify its appearance and voice according to their needs. The interview can consist of several parts, including a situational component that allows the candidate to demonstrate their ability to interact with clients or colleagues. An example of a situational setup for Porsche is available at this link (demo test): https://www.sweeeft.app/public/porsche			

2.1 Objectives of the project and context

The objective of the project is to develop an innovative solution that meets the requirements of recruitment processes through a robot capable of conducting natural conversations. This robot, easily accessible via a web application, aims to simplify and optimize these processes by intelligently adapting to user responses. The tool will be available 24/7, which suits candidates currently employed who cannot answer their phone at their current employer. The project will complement an existing part of sweeeft.ai's recruitment software. This existing part already allows for the evaluation of candidates after the interview through various processes that analyze the recording of the candidate's camera and microphone, particularly their vocabulary, gestures, tone of voice, etc.



Studio de paramétrage de l'avatar/robot

The product must be ready to be marketed by the end of 2024.

2.2 State of the reference market

2.2.1 The Vera Robot

Vera, from the company Stafory, is a virtual robot with a female appearance, marketed in 2018 to help large companies recruit. She can call candidates and conduct interviews over the phone or via Skype. Equipped with artificial intelligence, she can respond to certain questions or pretend to bounce back from a response by saying 'Brilliant!' before moving on to the next point. On the screen, she has a human and feminine appearance. Her lips move in sync with her voice, and she slightly turns her head to simulate interest. Vera has been endorsed by numerous companies, including PepsiCo, Ikea, and L'Oréal.

Robot Vera

2.2.2 Manpower

Zara is the name of the assistant used by consultants at Manpower. This virtual character, represented by an avatar, automates the preselection of candidates by asking them questions to assess their ability to fill the position. A link is sent to the candidate's smartphone or computer, connecting them to Zara. The questions are those that a recruiter would have asked over the phone during the initial contact with the candidate. The script is customized by the employer based on the profile sought. Therefore, the Zara avatar is not capable of improvising a conversation based on the candidate’s responses or answering questions unlike the robot Vera.

Stafory and Manaround have withdrawn their recruitment tools from the market.

2.2.3 Comparison with the current solution

The solution we are developing presents several notable differences from previously developed assistants. The responses of our new avatar use the latest language models such as GPT-4 to generate its answers. This allows recruiters to easily give instructions to the avatar while letting it adapt to the candidate's responses. The robot Vera was developed in 2018, at a time when major language models were not yet commercialized. Furthermore, Vera's responses are entirely scripted.

The physical aspect of the new avatar uses a video of a real person by overlaying lip movements. This enhances the ability for recruiters to create an avatar with well-defined roles during role-playing scenarios. On the other hand, the physical appearance of Vera is that of a synthetic animation from the 2000s, thus less natural and less adaptable.

It is also interesting to note that unlike previously developed products, the new avatar does not have a name as it is completely modular and therefore does not refer to any particular character. This is crucial when wanting to conduct scenarios with multiple different avatars.

2.3 Performance targets

We aim to achieve the following performances:

- The dialogue between the avatar and the candidate/client must be fluid;
- Response times should be short, between 2 and 3 seconds after the candidate finishes speaking, at maximum;
- The avatar's lip movements must be synchronized with the voice;
- The avatar must follow the order in which the questions were set during configuration to facilitate the subsequent evaluation of the candidate's responses;
- The avatar must be able to orchestrate the triggering of visual elements (button, multiple-choice quizzes, catalogs...) that corroborate the interview;
- The tool is made available to the candidate/client through a website;
- The robot must be able to adapt to the candidate's responses;
- The role, physical appearance, and voice of the avatar must be fully customizable by the recruiter.

The last two performances above are totally distinguishing compared to previously existing solutions, and the combination of all these performances makes it a unique solution in the market.

Moreover, the solution we are developing does not only involve creating job interviews but also situational exercises with different avatars playing various roles, which is why it's important to freely adjust their physical appearance, voice, and personality. For instance, the candidate will take on the role of a salesperson and face an avatar playing the role of a customer at Picard (demo test):
https://www.sweeeft.app/public/picard/?campaignId=23749
Another example would be a situational exercise for a job in a nursing home where the candidate will be assessed on their relational skills facing an avatar playing an elderly person and another avatar playing the son of the elderly person.

2.4 Innovation work performed

My work on the project involved prototyping the solution and developing technical components to achieve the desired performance. In this context, I worked on:

- Improving response time;
- Generating lip movements;
- Engineering the prompt;
- Developing streaming communication protocols;
- Developing the cloud infrastructure.

My work and some of their results are presented below.

2.4.1 Study of response time

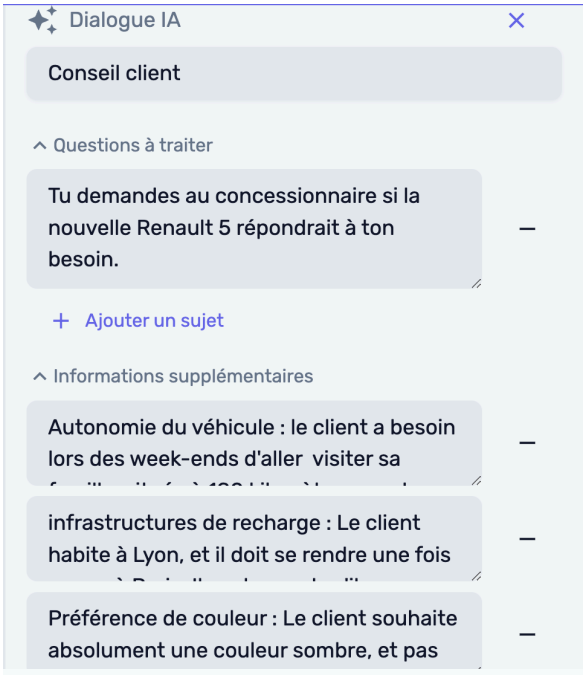
To assess the feasibility of the project, the first part of my work involved developing a web application that allows interaction with OpenAI through speech. The issue at hand is the response time; each time ChatGPT finishes a sentence, the voice synthesis is produced and broadcast to the user while the next sentence is being synthesized. If the user decides to add something, then the processing of the avatar's response is stopped in order to have a smooth conversation.

2.4.2 Study of lip movement generation

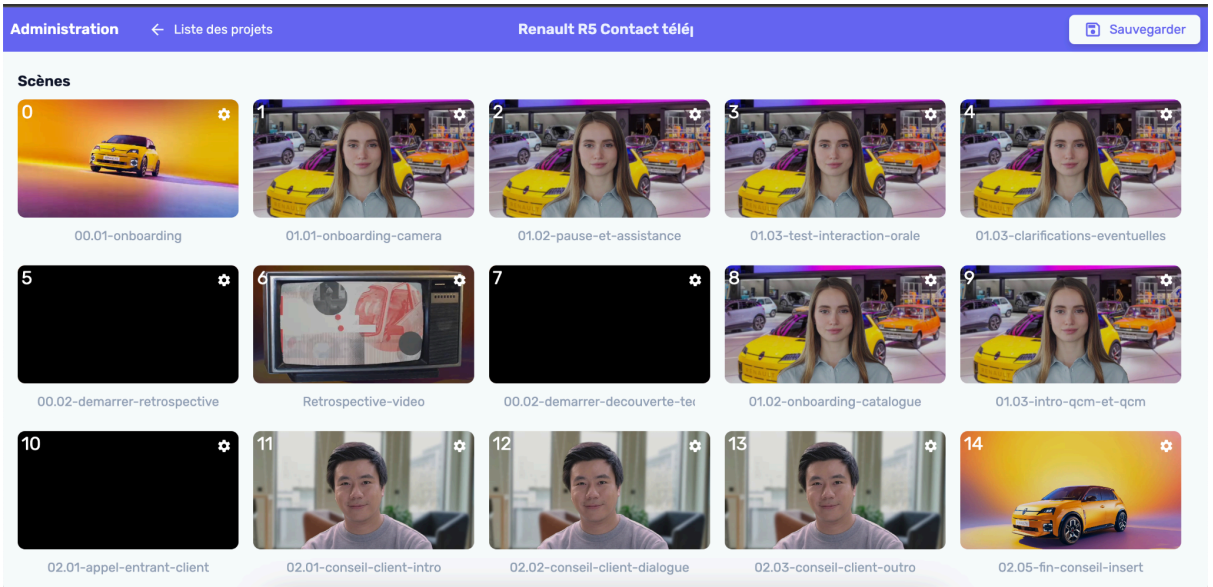
Once the avatar's voice is synthesized, the audio is input into a generative model for lip movement. The lip movement generation is produced in real time, meaning the time between each frame is sufficient to generate the lip movement. Currently, there is no API capable of returning a stream of images that animates lip movements in real time. Therefore, this part is handled on our own servers equipped with GPUs. This was made possible thanks to my research on the state of the art in the field of generative models. Since the academic field is often removed from the constraints of speed, I also had to perform optimization work. Only the part of the face where the mouth is located is generated and then superimposed on a base video representing the avatar. This base video can theoretically be chosen freely, but the quality is not always good, especially if the chosen video is far from the training set of the generative model. The evaluation of lip movement performance is qualitative.

2.4.3 Prompt engineering

Once the avatar is able to interact naturally with the user, my work involved allowing the customization of the role and conversation themes with the avatar. A prompt is created and input into OpenAI's API. This prompt consists of fixed parts, but also parts that the recruiter can modify based on their expectations through an interface. Here is an example in a scenario where the avatar will play the role of a customer at Renault:



The recruiter has the ability to segment the interview into several parts in order to control the timeline in which different avatars will intervene, as well as the timeline of the topics discussed and the visual elements (buttons, images...) that can be displayed.



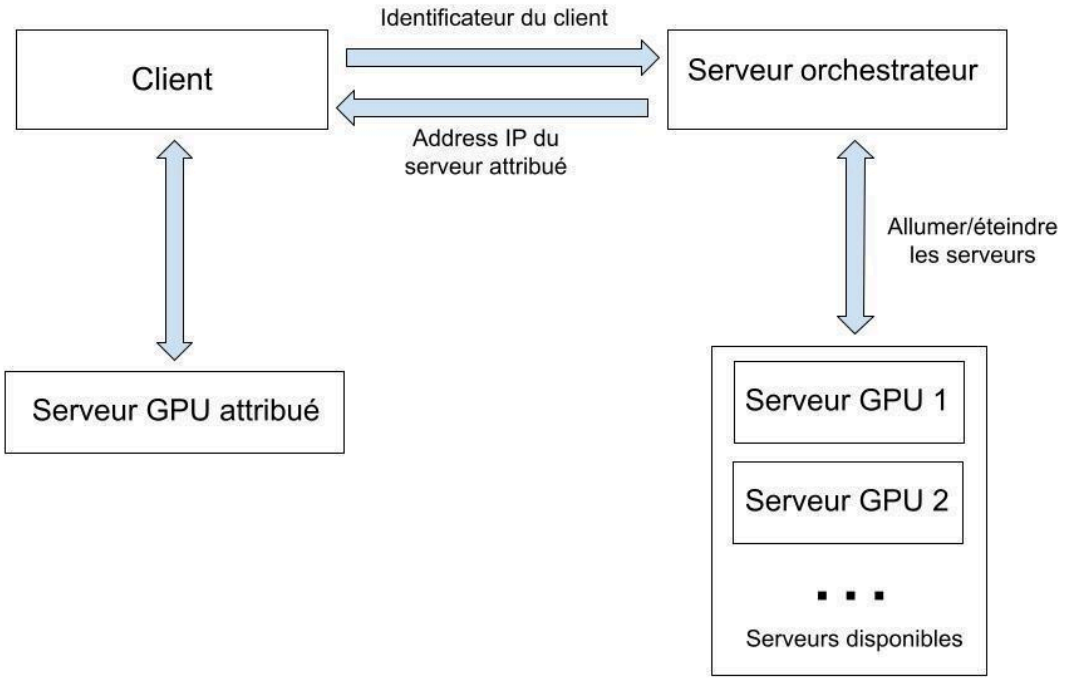
The prompt is segmented into several parts, and OpenAI will thus indicate the change of part through a tag, 'Part n:'. The need to segment the conversation also aims to target analyses for the automatic evaluation of the candidate's responses.

2.4.4 Development of streaming communication protocols

A part of my work focused on streaming communication protocols. The user's voice is recorded and sent in streaming to the server to be analyzed. The communication protocol used here is WebSocket. In return, the client receives a video stream representing the avatar, which requires a judicious choice of video compression to ensure good functioning even in low bandwidth conditions. This time, the communication protocol used is AIORTC.

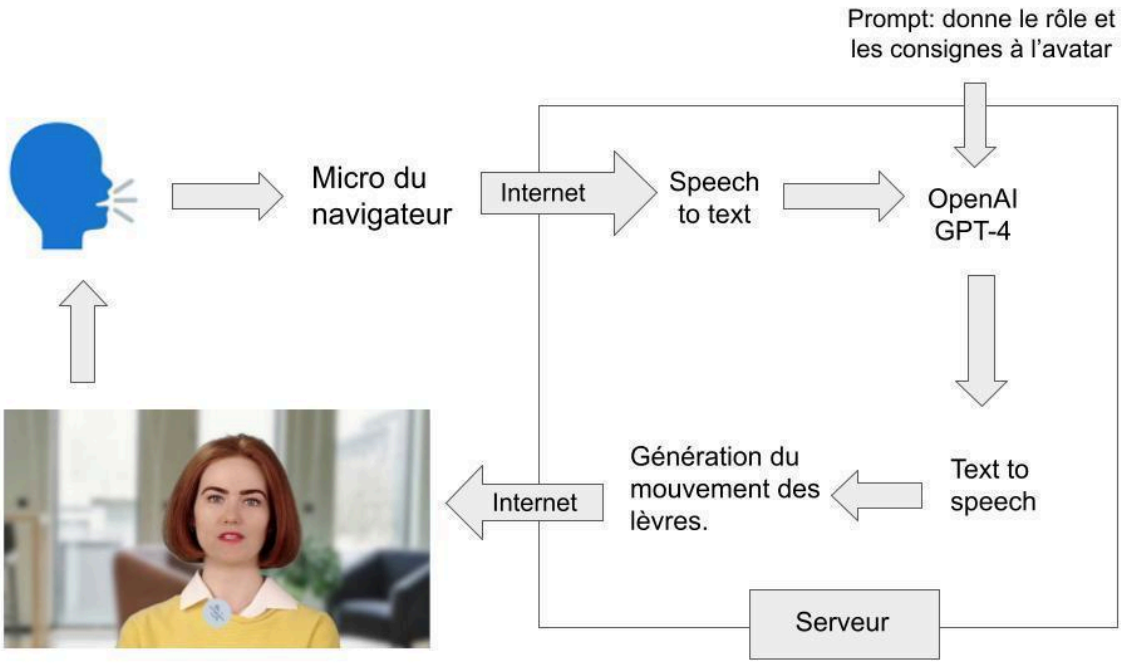
2.4.5 Development of cloud infrastructure

The final stage of my work involved setting up the cloud infrastructure to ensure that multiple users can simultaneously use the avatar. Each client utilizes a GPU server (NVIDIA T4) and a system manages the startup and shutdown of the servers based on the number of clients to optimize costs. When a client connects, a server (GPU) is powered on in advance to allow for a quick start. The diagram of the infrastructure is presented below.



Each response from the avatar requires four steps (as outlined below):

- Speech to text for transcribing the user's speech;
- A request to the OpenAI API;
- Text to speech for the avatar's voice synthesis;
- Lip movement generation.



2.4.6 Planned improvements

The avatar's response time (between 3 and 4 seconds) still limits the fluidity of the dialogue. Recent advancements by OpenAI in the field of oral interactions with GPT-4o show significant improvements in response time. Instead of segmenting each step (speech to text, OpenAI's response, text to speech…), the new GPT-4o model directly takes an audio stream as input and outputs an audio stream that synthesizes the avatar's voice. This “end-to-end” solution also allows the model to analyze the emotion in the user’s voice, thus better interpreting their response, where currently the speech-to-text reduces the information contained in the interaction.

Another possible improvement would be to allow the user to interrupt the avatar. This is not currently possible because it is difficult to filter out the avatar's voice that overlaps with the user's voice.

2.5 Location of work execution, specific materials and means implemented

The development took place in Lyon, with AWS supporting the hosting of the servers, and a second person from Sweeeft.ai worked on the user interface for configuring the avatar. In addition to OpenAI, several APIs are used. These include ElevenLabs for voice synthesis and Deepgram for speech to text. I worked on this project as a developer in artificial intelligence and cloud.

2.6 Total project cost and amount of company participation

The project, led by two developers, incorporates costs related to personnel and server use. The personnel cost is €58,000 per person per year. The project is expected to be fully operational after one year of development. The servers cost €0.75 per hour, amounting to an annual cost of €6,570. Additionally, there are costs associated with the use of APIs such as OpenAI and ElevenLabs, costing €6,000 and €7,200 per year, respectively. This brings the total annual cost to €135,000.

2.7 Innovation indicators

This project is not subject to grants or intellectual property protection.