Speech Translation System

Report by

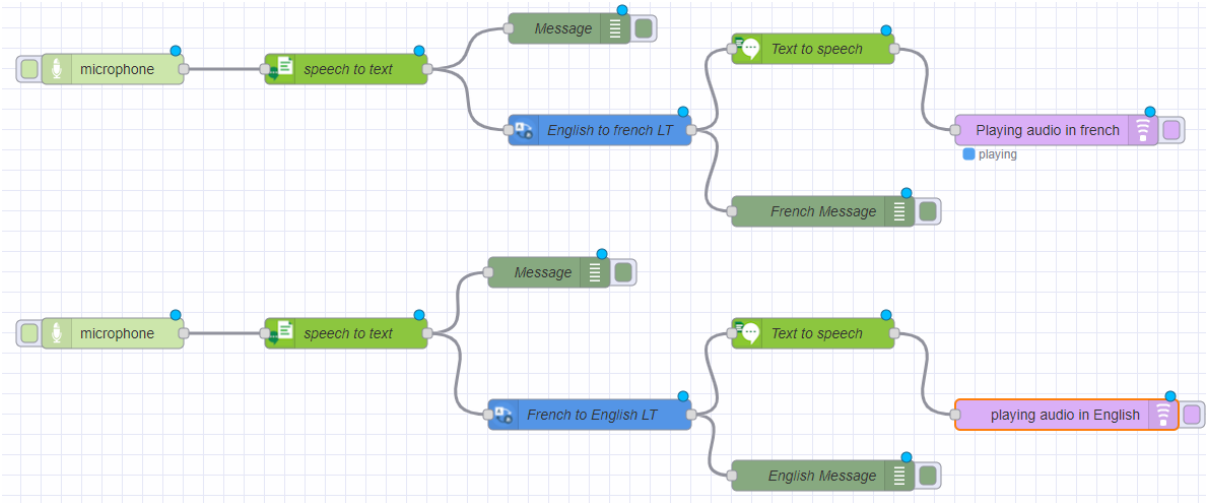
Ariane Shimirwa

ashimirw@andrew.cmu.edu

November 7th 2023

1. Creation of a speech-to-speech translation system

A. Final system Flow



B. Development of the system

To develop this system and get to the final system flow above, the following steps were followed.

The first thing we did was create an **account** on the **IBM cloud** and use the cloud feature code to receive the needed privileges from IBM. We then created the instances of the **Watson AI services** on IBM. We created the **speech-to-text**, **text-to-speech,** and the **Watson language translator**. These services can be found in the catalog on the IBM cloud. The most important thing was retrieving the API keys and the URLs for these services. Figure 1, 2, and 3 shows **the speech-to-text, language translator, and text-to-speech Watson AI service** we created and their credentials.

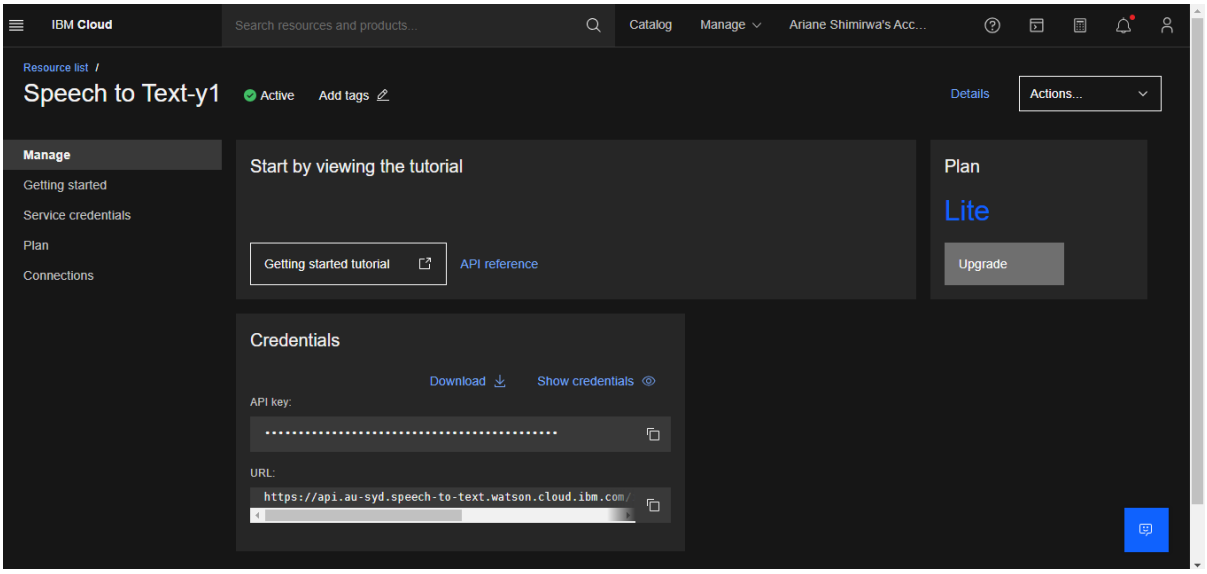


Fig 1. Credentials of Watson speech-to-text (STT) Service

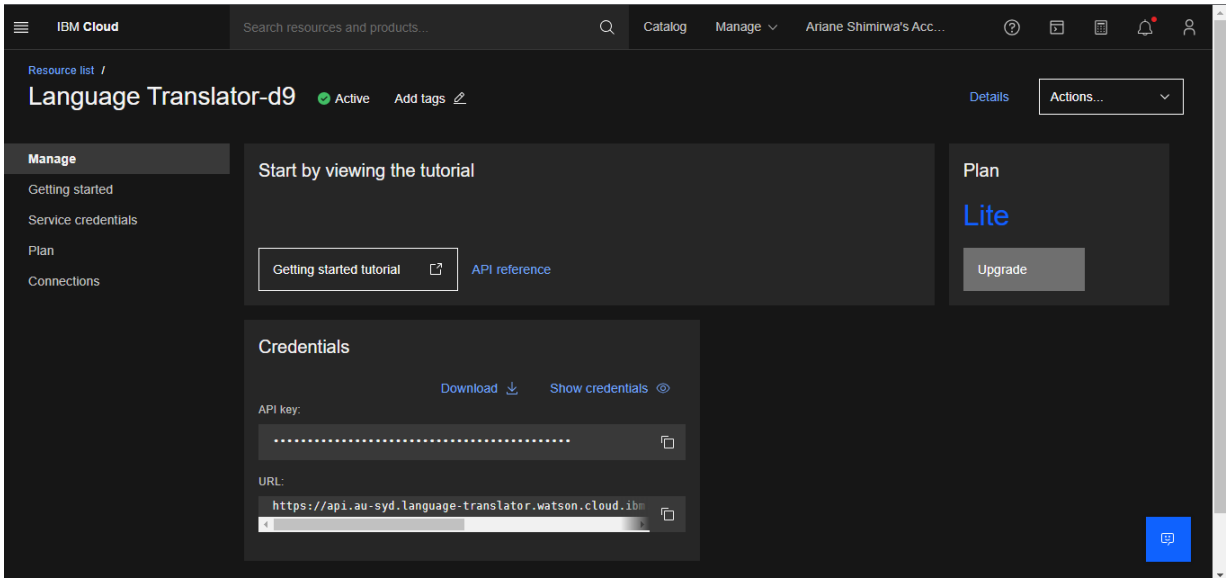


Fig 2. Credentials of Watson language translator (WLT) Service

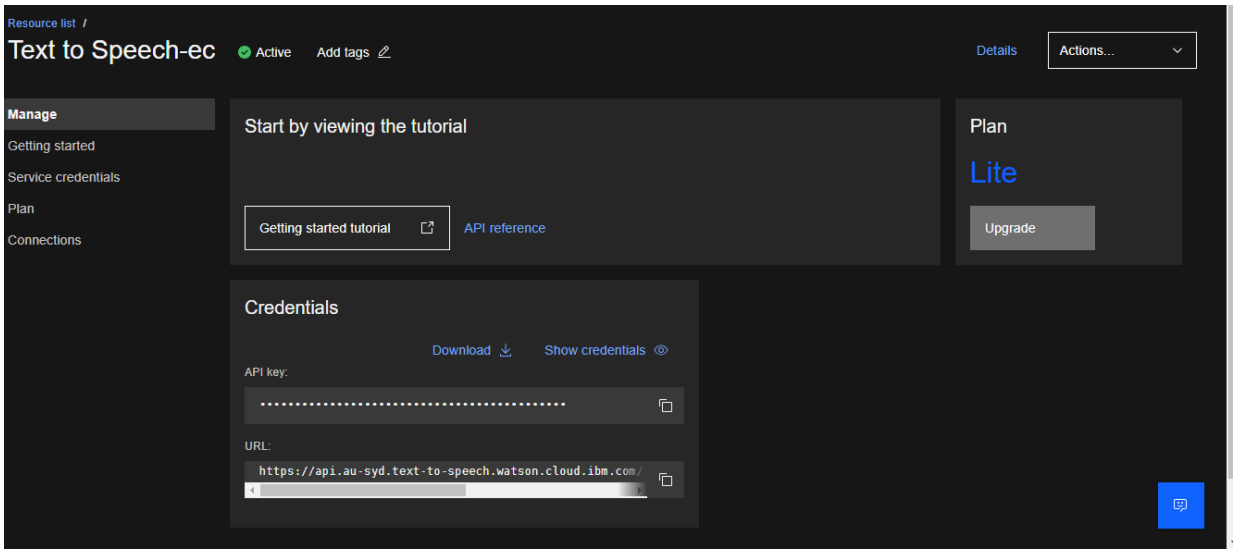


Fig 3. Credentials of Watson Text-to-Speech (TTS) Service

The next step was to create the **flowfuse** account. After creating an account, we created our application through the **Node-RED visual programming editor**.

We also installed some additional nodes before we started creating our application. The installed nodes are **node-red-contrib-browser- utils node,play audio, and node-red-node-watson node** to help us build the system.

How the system works

The final system flow shows a two-part flow designed for translating spoken language, with the first flow illustrating English-to-French translation and the second illustrating French-to-English translation.

English to French translation

The English-to-French translation flow starts with a microphone node, where the system captures audio of the user's spoken English words. This audio is then processed by a speech-to-text (STT) service, converting it into written text. The English text is then directed into an

English-to-French language translator service, where it is translated into French. The resulting French text is then fed into a text-to-speech service that transforms the written French text back into spoken audio. Finally, the system outputs the translated French audio through a playing audio node, completing the translation process and the user can hear it.

Figures 4, 5, and 6 detail the configurations of various nodes used in our speech translation setup of English to French, utilizing IBM's Watson AI services.

Figure 4 displays the speech-to-text node configuration, employing IBM's technology with the necessary API key and service endpoint URL, set to process English audio with default settings. Figure 5 outlines the language translator node's settings, with the API key and URL linked to our IBM Watson AI service instance, configured to translate from English to French in 'General' domain mode, with parameters activated (The source language is English and the target language is French) for the translation process. In Figure 6, the text-to-speech node is shown using the relevant API key and URL set to convert text into French speech specifically with the 'Renee' voice in WAV format and outputs the audio to msg.payload.

Finally, Figure 7 presents an example of the program in action during a test, effectively translating text from English to French and presenting the translated text in the debug console.

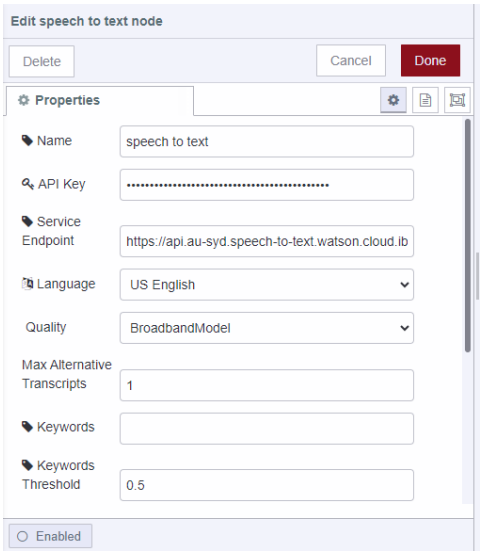


Fig 4. Speech-to-Text configurations

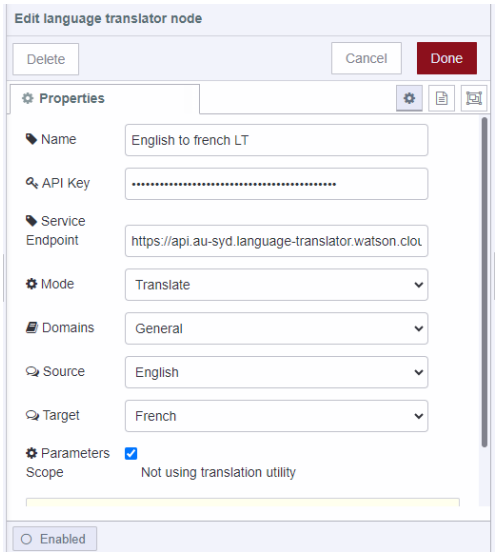


Fig 5. Language translator configurations

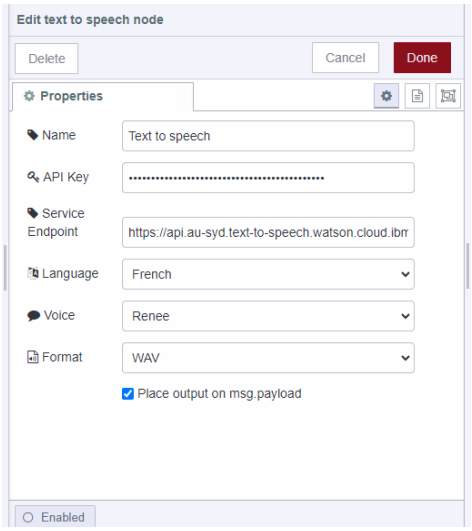


Fig 6. Text-to-Speech configurations

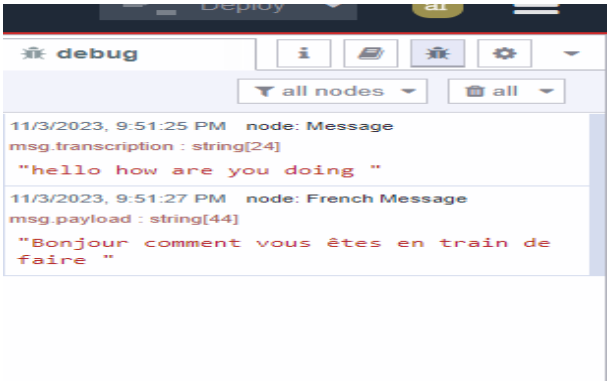


Fig 7. Example of English to french text translation

French to English translation flow

The French to English translation flow initiates with a microphone node, where the system captures the user's spoken French. This audio is then processed by a speech to text (STT) service, converting it into written text. The French text is directed into a French to English language translator service, where it is translated into English. The English translation is then fed into a text to speech service, which transforms the written English text back into spoken audio. Finally, the system outputs the translated English audio through a playing audio node, completing the translation process and the user can hear it.

Figures 8, 9, and 10 detail the configurations of various nodes used in our speech translation setup of French to English, utilizing IBM's Watson AI services.

Figure 8 displays the speech-to-text node configuration, employing IBM's technology with the necessary API key and service endpoint URL, set to process French audio with default settings. Figure 9 outlines the language translator node's settings, with the API key and URL linked to our IBM Watson AI service instance, configured to translate from French to English in 'General' domain mode, with parameters (the source language is French and the target language is English) activated for the translation process. In Figure 10, the text-to-speech node is shown, using the relevant API key and URL, set to convert text into English speech, specifically with the 'AllisonExpressive' voice in WAV format, and outputs the audio to msg.payload.

Lastly, Figure 11 presents an example of the program in action during a test, effectively translating text from French to English and presenting the translated text in the debug console.

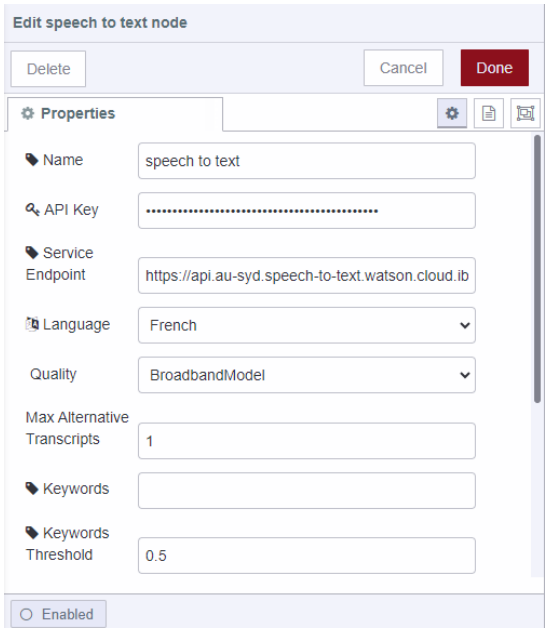


Fig 8. Speech-to-Text configurations

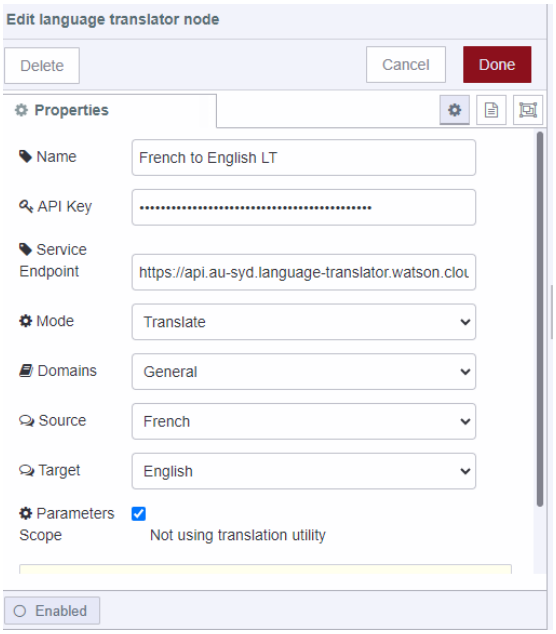


Fig 9. Language translator configurations

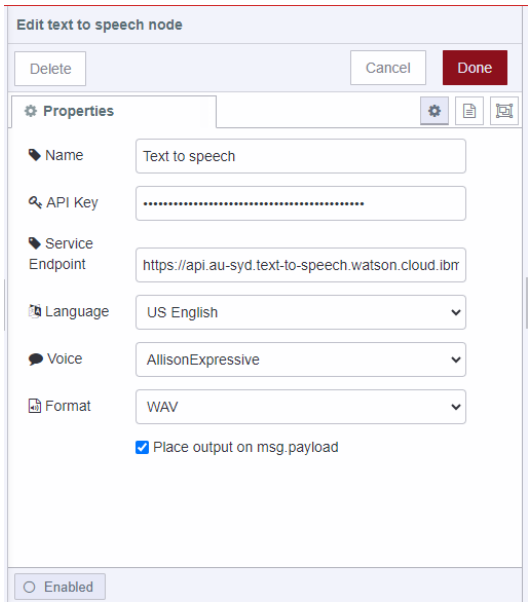


Fig 10. Text-to-Speech configurations

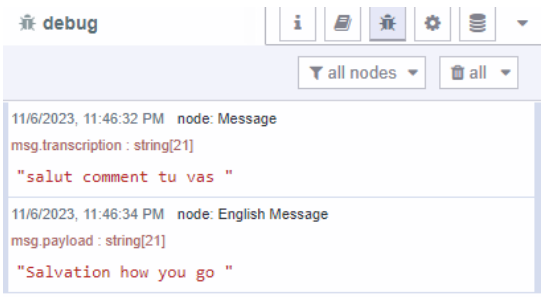


Fig 11. Example of English to french text translation

The function of each function

1. **Nodes**
   - **Microphone node:** A microphone acts as the initial point of contact with the user capturing the user's audio (spoken language) and transforming the sound waves from the user's voice into electrical signals that can then be passed into the speech-to-text service.

   - **Play audio node:** The play audio node here takes the translated text to speech and translates it into an audible form by playing the final translated result.

2. **Technology**
   - **Automatic Speech Recognition:** this is the process that occurs when the microphone captures spoken words and converts them into text through the 'speech to text' node. This enables the system to understand and transcribe spoken language into written form.
   - **Machine translation:** this is the process that applies a translation model specifically using Natural Language Processing (NLP) models to accurately translate the text from one language to another. It interprets the context and meaning behind the original text to produce a coherent and equivalent text translation in the targeted language.
   - **Speech Synthesizer:** this process takes the translated text and converts it back into an audible speech. This technology uses complex algorithms to produce natural-sounding voice output from written text, allowing the receiver to hear the message in the language they understand.

3. **Analytics**
   - **Speech-to-Text (STT):** It captures and converts the verbally spoken language into written form(text), enabling seamless interaction with devices by simply speaking, and it creates a text-based record.
   - **Text-to-Speech (TTS):** It captures speech in the form of a text and converts it into an audible speech serving as an audible bridge between written content and spoken language.
   - **Language translator:** It analyzes text in one language and converts it into another language. It is an analytical process that uses NLP models to understand the context, grammar, semantics, and syntax of one language and provide accurate translations into another one.

2. Video Creation of two people holding a conversation using this translation system

The selected languages for the conversations are English and French.

A. **Scripted conversation (Two people making introductions):**

David : Bonjour

Ariane : Hello

David : Comment ça va ?

Ariane: It is going very well, thank you. And how about you?

David: Ça va, merci. Comment vous appelez-vous?

Ariane: My name is Ariane, and you? What is your name?

David: Je m'appelle David.

Ariane: Nice to meet you, David.

David: Enchanté(e), Ariane.

Ariane: And where are you from David?

David : Je suis ingénieur au Kigali.

Ariane: That is amazing David. I am a pilot in RwandaAir.

David : C'est charmant. Ça doit être formidable de visiter différents pays et de rencontrer de nouvelles personnes.

Ariane: That is great David. It was nice to meet you.

David : C'est un plaisir de te recontrer aussi Ariane.

Link to the Video: Video 1

B. **The random conversation subject is School Holiday Break**

This conversation is about two people talking about what they did in their school holiday break. Since it is not scripted we did not provide a script for it but you can refer to the link below to view it.

Link to Video: Video 2

C. Comparative analysis of the two conversations

General analysis

The system exhibits a high level of accuracy when translating from English to French; however, it tends to perform less effectively when translating from French to English. The translation process can be sluggish at times, resulting in slower output. Moreover, occasional inaccuracies in translations may lead to misunderstandings with the intended recipient. In some cases, the presence of a poor accent can contribute to unintended translations. Additionally, when handling lengthy audio content, the system's translation process can become time-consuming. Speaking slowly and with a clear, loud voice can contribute to improved accuracy in the translation process. This allows the system to better capture and interpret the spoken language, resulting in more precise translations.

Comparative Analysis

Scripted conversation

- **Accuracy is high:** The accuracy of translation in a scripted conversation was generally good, particularly when examining most of our dialogue. For instance, in Figure 12 below, it's evident that the system accurately translated the conversation where two individuals introduced themselves. Nevertheless, while most of the translation was correct, the system encountered challenges in its speech-to-text conversion, particularly concerning proper nouns. An illustrative example is its conversion of the name "Ariane" to "Ken" which highlights a clear instance of the system's difficulty in correctly recognizing and transcribing proper nouns.

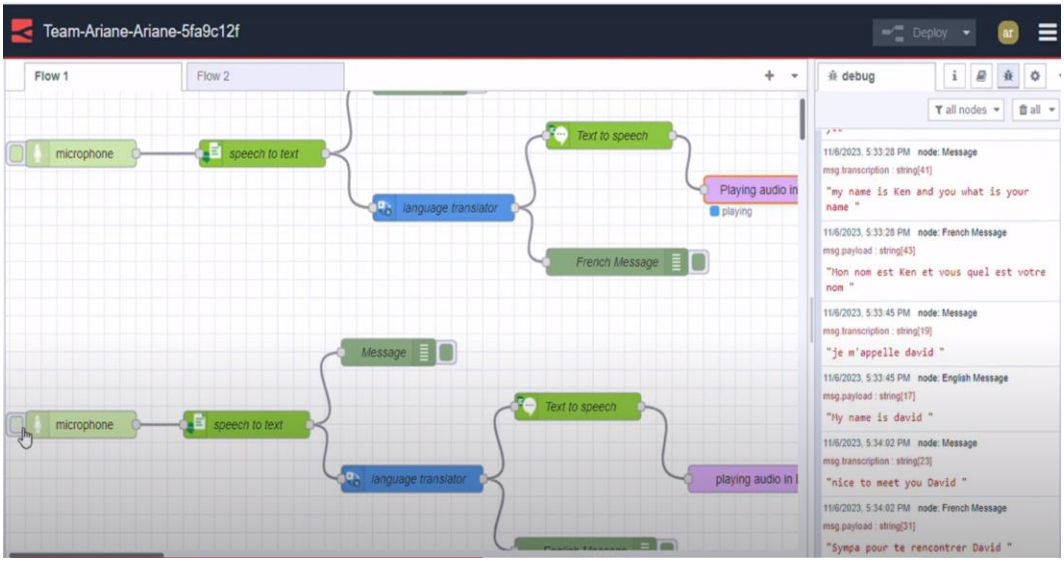


Fig 12. High accuracy example in scripted conversation

- **Translation is fast:** Our analysis of the scripted conversation also highlighted the system's remarkable speed in delivering translation results. As evidenced in Video 1, the system efficiently provided translations from one language to another. We assume that this speed is related to the nature of scripted conversations, where participants typically do not require extra time to formulate their speech. This streamlined process

makes it easier for the language translation system to perform its task promptly and effectively.

Random conversation

- **Accuracy is low:** In our analysis of testing the language translation system using unscripted conversations, we found that the system did not consistently produce accurate translations. These natural dialogues often involve pauses and the use of filler words as speakers gather their thoughts. Unfortunately, the system sometimes misconstrued these filler words such as "ah" as meaningful content, leading to the inclusion of extra words that were not part of the original conversation. This resulted in less precise and sometimes disorganized translations as you can see in the Figure 13 below.

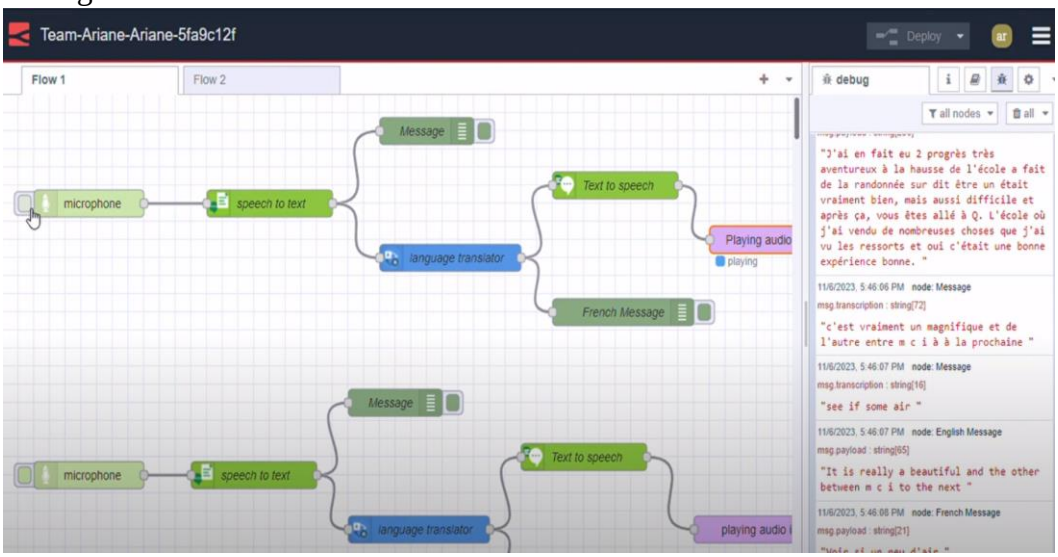


Fig 13. Example of low accuracy in random conversation

- **Translation is slow:** Translation speed is reduced in random conversations, largely due to the spontaneous nature of the dialogues. In these situations, participants often take additional time to carefully formulate their words. Consequently, the system also requires slightly more time to deliver translation results, as illustrated in Video 2. When participants engage in more thoughtful speech, the system's translation process naturally becomes more time-consuming, aligning with the pace of the conversation.

In summary, our observations indicate a notable disparity in user experience between two distinct conversation types. Scripted conversations offer a significantly superior experience characterized by higher accuracy and fast translations. In contrast, random conversations provide lower accuracy and slower translation, highlighting the importance of context and conversational dynamics in shaping the performance of language translation systems. Therefore, for better user experience and performance, we recommend IBM to enhance this system by improving context awareness, handling filler words, recognizing proper nouns, adding more diverse training data, and implementing adaptive algorithms to cater to different conversational dynamics.

3. Testing the machine translation system by doing Text-to-Text translation

English to French text and French to English System Flow

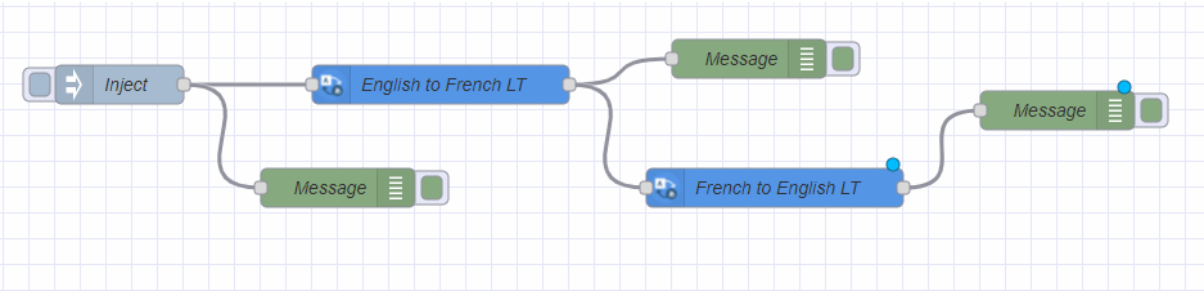


Fig 14. English to French – French to English System Flow

Figure 14 shows a flow that starts with the injection node that takes in text paragraphs and outputs the English text message to the console before being processed in the English-to-French language translator to be translated to French. When it is translated to French, it outputs the translated message, and the translated message gets translated back again into English using the French-to-English language translator.

The random paragraphs used in the injection node:

He watched as the young man tried to impress everyone in the room with his intelligence. There was no doubt that he was smart. The fact that he was more intelligent than anyone else in the room could have been easily deduced, but nobody was really paying any attention due to the fact that it was also obvious that the young man only cared about his intelligence.

He slowly poured the drink over a large chunk of ice he has especially chiseled off a larger block. He didn't particularly like his drinks cold, but he knew that the drama of chiseling the ice and then pouring a drink over it looked far more impressive than how he actually liked it. It was all about image and he'd managed to perfect the image that he wanted to project.

A. Results

Figure 15 shows the results from the debug console of the injected paragraphs translated into French. Figure 16 shows the results from the debug console of the translated French text being translated back into English.

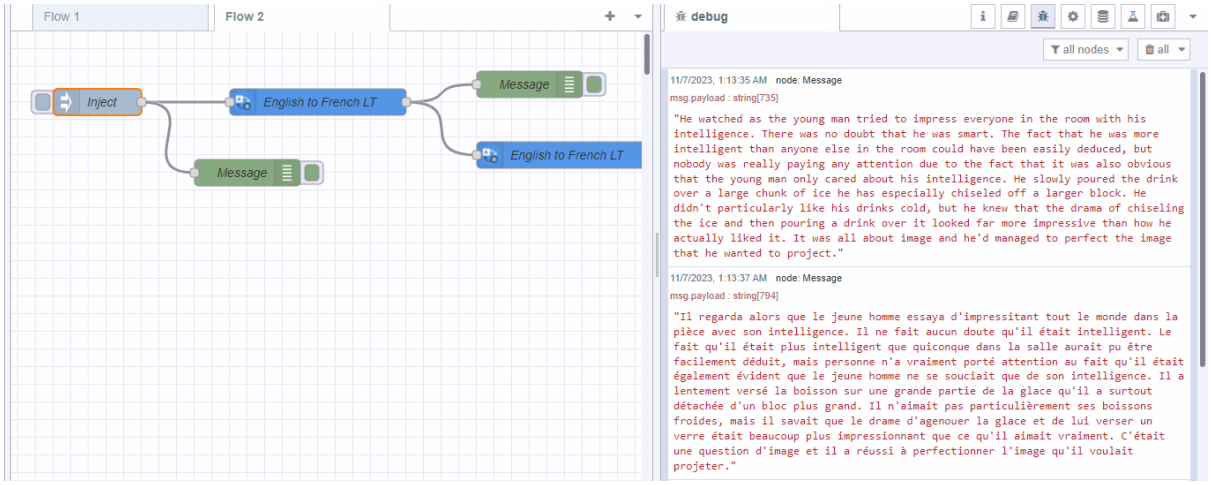


Fig15. Text translation from English to French

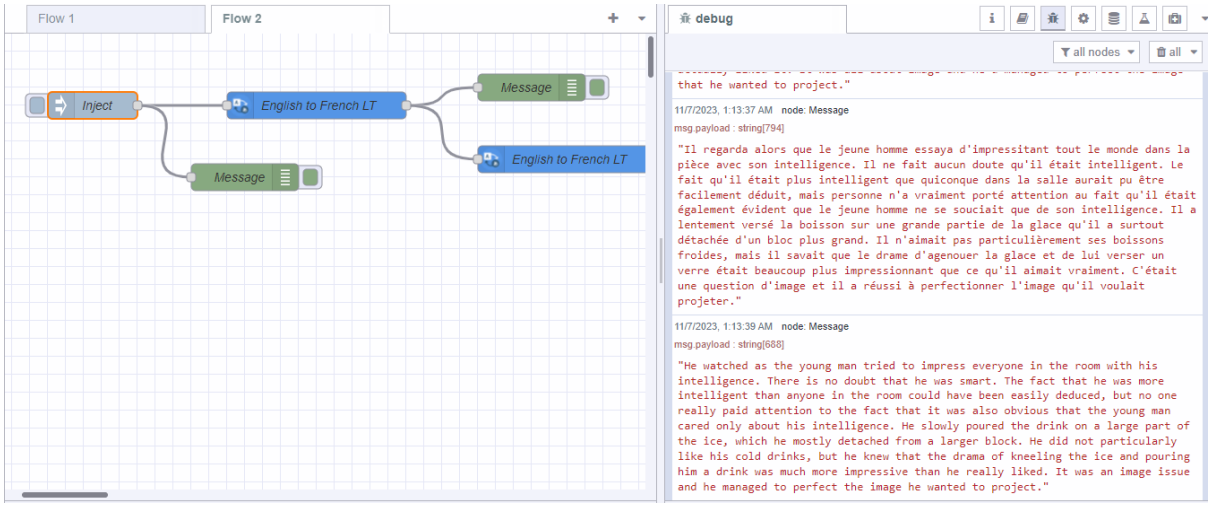


Fig16. Text translation from French to English

Comparing the original text and the final English text from the system, they are almost similar. There were some minor changes in the resulted English text from French where it paraphrased some of the difficult terms that were in the original text. For example, On line 6 from the original text**, "chilesed off a larger block"** was paraphrased to **"detached from a larger block"** in the converted back to English paragraph and the last sentence which was **" It was all about image and he'd managed to perfect the image that he wanted to project"** in the original text was paraphrased to **"It was an image issue and he managed to perfect the image he wanted to project"** in the converted back to English paragraph which is almost the same but a bit paraphrased.

B. Commenting on the results

To conclude text-to-text speech translation seems to be more accurate using English to French and French to English since the returned results after translation conveys the same message as the original translation. The converted message is a bit paraphrased, but it does not change the original message.

Uses of speech-to-speech translation system in Africa

1. **Education**: 60% of youth between the ages of 15 and 17 are not in school [1] and to add more to this Africa lacks enough learning material for students [2]. Using the language translation system to translate material from other continents into the local languages of Africans can allow students in Africa to have access to a wider pool of educational material which can be a big step towards solving the education challenge in the continent.
2. **Healthcare**: Health information could be made more accessible to patients who speak various local languages, improving understanding and compliance. There is a lack of sufficient and quality healthcare-related data in Africa [3]. A language translation application that can translate healthcare information into African local languages can help Africans get good quality healthcare information.
3. **Commerce**: Traders and entrepreneurs could communicate with a broader market, both within and outside Africa. According to Venture Africa, the Rwandan private sector was losing business due to language barrier [4]. A language translation system would help African businesses solve the language barrier issue and be able to compete not only on a regional level but also globally.
4. **Government Services**: It would enable governments to deliver services and disseminate information in all official and local languages, ensuring inclusivity. With over 2000 living languages on the continent [5] it is very hard for governments to effectively deliver information to all citizens which leads to reliance on a popular language that usually is not understood by the whole population. Implementing a translation system in this case can help remove reliance on popular languages and allow governments to communicate effectively to their citizens in their local languages.
5. **Tourism:** A study done by the Rwanda Development Board (RDB) showed that 85,730 workers in the hospitality sector were not able to be fluent in foreign languages [6]. Language barriers can lead to poor customer service leading to customer dissatisfaction and eventually losses. Building a language translation application that can help tourists and local employees to communicate with each other can help boost the tourism industry.

Challenges to be Solved

1. **Accents and Idioms:** Africa is home to over 2000 living languages [5]. Each language has a wide variety of accents and regional idioms that could complicate speech recognition and translation. This would be a major challenge that would need to be solved to successfully implement this solution in Africa.
2. **Data Availability:** There is a lack of data in Africa, and where it is available it often of poor quality [7]. This can affect implementing language translation models that rely on good and sufficient data for training. Solving the issue of insufficient and bad-quality data specifically the one's about local languages would be a major milestone in implementing this solution on the continent.
3. **Technological Infrastructure:** Reliable internet access and electricity are still not universal across the continent, which could limit the use of online translation systems

[8]. Providing cheap and reliable internet in Africa is one of the issues that need to be addressed for this solution to work.

4. **Scarce IT knowledge to implement the solution:** The effective use of such technology requires a certain level of education and technological literacy, which is a rare resource across the continent[9]. Tackling this issue can bridge the gap towards implementing this solution in Africa.
5. **Cost:** 54.5% of the African population is under severe poverty [10]**.** The development, deployment, and maintenance of such systems can be expensive, and making them affordable and accessible is a significant challenge that needs to be addressed in order for this solution to work in Africa.

References

[1] ‘Why education remains a challenge in Africa – DW – 01/24/2022’. Accessed: Nov. 07, 2023. [Online]. Available: https://www.dw.com/en/africa-right-to-education-remains-a-challenge/a-60518000

[2] ‘Inadequate school and teaching resources challenge education in Sub-Saharan Africa - World | ReliefWeb’. Accessed: Nov. 07, 2023. [Online]. Available: https://reliefweb.int/report/world/inadequate-school-and-teaching-resources-challenge-education-sub-saharan-africa

[3] S. M. Musa *et al.*, ‘Paucity of Health Data in Africa: An Obstacle to Digital Health Implementation and Evidence-Based Practice’, *Public Health Rev.*, vol. 44, p. 1605821, 2023, doi: 10.3389/phrs.2023.1605821.

[4] N. Aderibigbe, ‘Language Barrier Hurting Rwanda’s Private Sector’, Ventures Africa. Accessed: Nov. 07, 2023. [Online]. Available: https://venturesafrica.com/language-barrier-hurting-rwandas-private-sector/

[5] ‘Africa: number of living languages by country 2022’, Statista. Accessed: Sep. 26, 2023. [Online]. Available: https://www.statista.com/statistics/1280625/number-of-living-languages-in-africa-by-country/

[6] ‘Language barrier for staff stifles hospitality industry’, The East African. Accessed: Nov. 07, 2023. [Online]. Available: https://www.theeastafrican.co.ke/tea/rwanda-today/language-barrier-for-staff-stifles-hospitality-industry--1314138

[7] A. Kinyondo and R. Pelizzo, ‘Poor Quality of Data in Africa: What Are the Issues?’, *Polit. Policy*, vol. 46, no. 6, pp. 851–877, 2018, doi: 10.1111/polp.12277.

[8] ‘Impact of the Internet in Africa’, A community connecting 1 billion people to the internet. Accessed: Oct. 31, 2023. [Online]. Available: https://unconnected.org/blog/impact-of-the-internet-in-africa-2021

[9] ‘Five unique cybersecurity challenges in Africa’. Accessed: Oct. 31, 2023. [Online]. Available: https://www.africa.engineering.cmu.edu/news/2023/08/23-cybersecurity.html

[10] ‘Africa needs to curb poverty and social inequality to meet development goals | Events | United Nations Economic Commission for Africa’. Accessed: Nov. 07, 2023. [Online]. Available: https://www.uneca.org/eca-events/stories/africa-needs-curb-poverty-and-social-inequality-meet-development-goals