Term Project / Cmpe48a

Berkay Demirtaş
Alper Canberk Balcı
Department of Computer Engineering
Bogazici University
Istanbul Turkey
alper.balci@boun.edu.tr
berkay.demirtas@boun.edu.tr

January 17, 2022

1 Introduction

There are many cloud service providers in the market and one of the most popular services that this companies provide is cloud-based AI service (AIAAS). Many different products such as object detection, sentiment analysis and translation tools are provided by companies. Market size of AIAAS is around 63 billion dollars and it is expected to grow by 14,7 billion in next 5 years. Evaluating the performance of the services in such a huge industry is an important issue. In this document, we will evaluate 3 AI APIs provided by Google Cloud Platform (gcp). These products are Speech-To-Text, Translation and Media Translation APIs. Speech-To-Text API is designed to transcribe speech audio to captions, Translation API, translates texts between two languages and Media Translation API directly translates speech audio to the captions in another language which means it does both speech to text operation and translation operation. We tried to convert a speech to a text in another language by using 2 different paths. One of them is using Speech-To-Text and Translation APIs and another one is using Media Translation API directly.

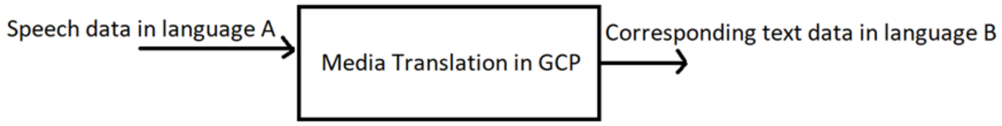

Speech data in language A
Media Translation in GCP
Corresponding text data in language B

Figure 1.1: Path 1

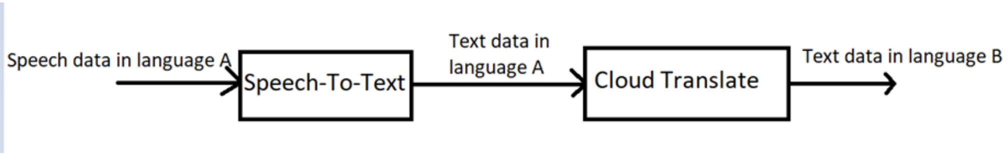


Figure 1.2: Path 2

By using the time required to make this calls and results of this calls, we analyzed the performances of this paths and compared this 2 paths. We also compared the costs for this paths for the same input data.

2 Related Work

3.1. Usage of related APIs in GCP

Before comparing 2 paths described in introduction section, we needed to learn what is the pricing for this products and how to make necessary calls for this APIs using Python.

3.1.1. Speech-To-Text API is designed to transcribe speech audio to captions.This API has 60-minute free tier and after 60 minutes, it charges 0.006 dollars per 15 seconds. Lenghts of the audio file is rounded up to 15 seconds. We used a tutorial and GCP documentation of this API to learn how to use this service.
There are 4 different ways to use this API. They are using gcloud tool from, command line, using command line, using Speech-To-Text UI and using client libraries.

We used client libraries in python and we will explain the details of this way.

- Install package : pip install –upgrade google-cloud-speech
- Create client object by using SpeechClient function of , create audio object from your file using RecognitionAudio funtion
- Create configuration object. To create this object, RecognitionConfig function is called. This function can take 19 different parameters but the parameters we used are sample_rate_hertz which is the sample rate of your record, enable_automatic_punctuation enables the punctuation in the resulting text and language_code is language of the record. Language code for English is 'en-US' . There is also a "model" parameter for this configuration. You can select suitable model according to your audio type (Video, phone call etc). We used default model which is not specialized for any specific type of audio.

3.1.2. Translator API is designed to translate a text from one language to another.This API translates 500,000 characters as free tier and after that 1 million characters cost 20 dollars. [**?**]
We used this API by using python client library.

- Install package : pip install –upgrade google-cloud-translate

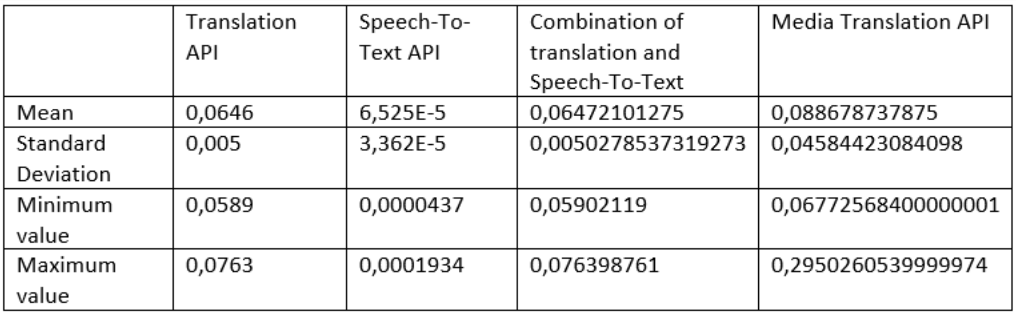
- Create client object by using Client() function of translate package, and send the input text by using translate() function of client object.You should also specify the target language as parameter. (It understands source language)

3.1.3. Media translator API is used for translating a speech record to text in another language. We used this API by using python client library and this is how we used it:

- Install package : pip install google-cloud-media-translation
- Create media translation client and configurations with TranslateSpeechConfig() function. This function takes 4 paramenters for default version of Media Translation Api (there are some spesific versions like for voice calls etc.). Those parameters are audio encoding, source language code ,target language code, sample rate hertz. We used English as source language and German as target language. Sample rate of our records is 48000 Hz.
- Create a request list by reading the voice files chunk by chunk. One chunk is 4096 bytes and for each chunk you should create a request with mediatranslation.StreamingTranslateSpeechRequest. After you create all the requests, call streaming_translate_speech() function of the client. This function accepts request list as parameter.
- You can get the results by traversing over text_translation_result field of the response.

3 Performance Evaluation

We used 24 different voice records with 48000 Hz sample rate. Average length of this records is 16.3 seconds. Here is the statistical values for API calls and 2 different path.



	Translation API	Speech-To-Text API	Combination of translation and Speech-To-Text	Media Translation API
Mean	0,0646	6,525E-5	0,06472101275	0,088678737875
Standard Deviation	0,005	3,362E-5	0,0050278537319273	0,04584423084098
Minimum value	0,0589	0,0000437	0,05902119	0,06772568400000001
Maximum value	0,0763	0,0001934	0,076398761	0,2950260539999974

Figure 3.1: Statistical values for response times of APIs and 2 different paths

When we check the mean value for path2 (Combination of translation and speech-to-text API) and path1 (Media Translation API), we see that mean response time of path1 is almost 37.5 percent higher than path2. This means on average, translating a speech in language A to text in language B is faster if we use path2.

We can see that standard deviation is larger for path1 because Media Translation spends more time for translating for longer speeches as it performs partial translations.

We were expecting mean response time of path2 to be larger than path1 because path2 contains 2 API calls and path1 contains 1 API call which means path2 contains more networking operations than path1. However, the results are showing the opposite. Therefore, we can conclude that network operation is a small portion of the response time of the API call. This can be proved further by the fact that both average response time of Speech-To-Text API and Translation API contains time for network and main operation, yet time for Translation API is way larger than Speech-To-Text API and networking time can't be larger than response time of Speech-To-Text API. Therefore, only small portion of average response time of these APIs is caused by network operations.

You can see the results of some API calls in below (figure 4.2).

4th sentence has been converted from speech to text successfully and translated correctly in the combined path. However, Media translation did not "hear" half of the speech and did not translate it because of that.

In the 5th sentence, Speech-to-Text and Media Translation APIs heard "There were paper..." as "Diverter paper, ...". Also, they translated the second part of the speech incorrectly. There shouldn't be "selbst" in this sentence in German.

Sometimes, APIs don't perform well, as examplified above.

We can argue that Media Translation is specialized in providing services for live streams. We don't know what it does to the speech when converting it to text, but we observed that sometimes Media Translation cannot "hear" the sentence completely. Also, it tries to translate the text in a different way. It translates the text (that it derived from the speech input) several times, performing partial translates. At the end it gives the final translation, and the result is almost always the same with the Translate API. However, Speech-to-Text API does a better job in catching the whole speech and converting it to text.

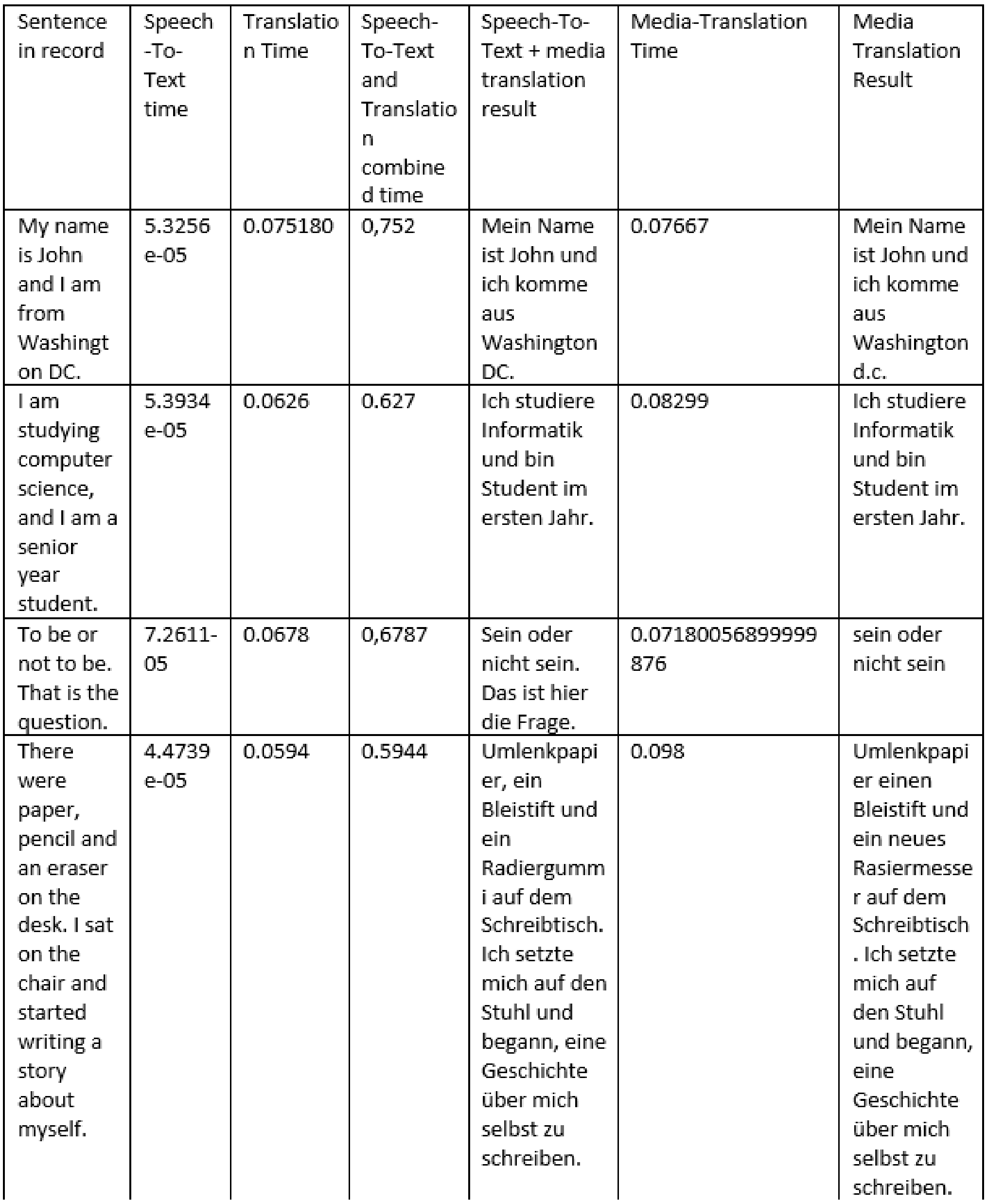
In addition, Media Translation works slower due to partial translations, lowering its performance further. For longer speechs, differences between performance of the paths is more significant again because of partial translations in Media Translation API.

When we compare the costs of this 2 paths money-wise, results are as follows:
Cost for Media Translation:

F(m) = $0.068 * (m - 60), 60 < m < 1,000,000$
$F(20,000)$ =1,355.92

Cost for Speech-to-Text and Translation together:

F(m) = (0.016+0.015) *(m-60) = $0.031 * (m-60), 60 < m < 1,000,000 (DataLoggingopt-in)$
$F(m) = (0.024 + 0.015) * (m - 60)$ =0.039*(m-60), 60<m<1,000,000 (without)
F(m) = $0.035 * (m - 60), 60 < m < 1,000,000 (average)$
$F(20,000)$ =697.9



Sentence in record	Speech -To-Text time	Translation Time	Speech-To-Text and Translation combined time	Speech-To-Text + media translation result	Media-Translation Time	Media Translation Result
My name is John and I am from Washington DC.	5.3256 e-05	0.075180	0,752	Mein Name ist John und ich komme aus Washington DC.	0.07667	Mein Name ist John und ich komme aus Washington d.c.
I am studying computer science, and I am a senior year student.	5.3934 e-05	0.0626	0.627	Ich studiere Informatik und bin Student im ersten Jahr.	0.08299	Ich studiere Informatik und bin Student im ersten Jahr.
To be or not to be. That is the question.	7.2611-05	0.0678	0,6787	Sein oder nicht sein. Das ist hier die Frage.	0.07180056899999876	sein oder nicht sein
There were paper, pencil and an eraser on the desk. I sat on the chair and started writing a story about myself.	4.4739 e-05	0.0594	0.5944	Umlenkpapier, ein Bleistift und ein Radiergummi auf dem Schreibtisch. Ich setzte mich auf den Stuhl und begann, eine Geschichte über mich selbst zu schreiben.	0.098	Umlenkpapier einen Bleistift und ein neues Rasiermesser auf dem Schreibtisch . Ich setzte mich auf den Stuhl und begann, eine Geschichte über mich selbst zu schreiben.

Figure 3.2: Some example translations by using 2 different path

4 Conclusion

Our experiment and analyses confirmed that using path2 gives faster and better quality results in terms of translations, while being cheaper than path1. Media Translation API does partial translation operations causing significant reduction in speed. Considering the cost and performance together, it even becomes less viable as an option. For now, combination of Speech-to-Text and Translation APIs seems significantly more powerful than Media Translation API except for live streams.

5 References

https://cloud.google.com/translate/docs/?_ga=2.98348632.-676760100.1637677409
https://cloud.google.com/translate/pricing
https://cloud.google.com/media-translation
https://cloud.google.com/speech-to-text/docs/basics
https://cloud.google.com/translate/media/docs/libraries/client-libraries
https://www.youtube.com/watch?v=lKra6E_tp5U

You can check the detailed work done from via :
https://drive.google.com/drive/folders/1nK1IQhPt6ipJ-47MWH_0kfAcJiopNCDx?usp=sharing