

Comparative Study Of Cloud Speech Recognition

CIS4010 Cloud Computing Team Project Report

Tyler Green, and Loui Zibdawi

School of Computer Science, University of Guelph, Guelph, Ontario, Canada
{tgreen10, lzibdawi}@uoguelph.ca

Keywords: AWS, Azure, Speech Recognition

Abstract: In this report we analyze two popular speech-to-text cloud services, Amazon Transcribe and Azure Cognitive Services: Speech. In order to test these services' ability to handle difficult speech, we measure their performance against audio clips of individuals with strong Irish and Canadian accents. From these tests we get measures of accuracy on a basis of word error rate (WER for short), which is the percentage of errors (Including incorrectly transcribed words, incorrectly added words and missed words) in the text that is produced. On top of individual accuracy ratings out of 10 for Irish and Canadian accents, we also give these services ratings on cost and implementation. This results in an point-based system where 40 is the highest possible score.

1 INTRODUCTION

This project was motivated by hearing about countless stories of people with accents other than a US/Canadian accent struggling with speech recognition services such as Cortana, Alexa, and Siri. We wanted to find out which service provider offered the best service to be able to allow other developers the opportunity to use the most accessible voice recognition service. We used a score based comparative approach. Each service would be implemented in python and then based on a number of categories each with their own score. At the end the scores would be totaled and the better service would be determined by the higher score. The accuracy measure chosen was word error rate or WER for short. This was chosen because it accurately captures all the mistakes accounting for additions, deletions, and incorrect transcriptions. With WER the lower the score the better the accuracy.

1.1 Project Statement

We aim to help identify the best Speech To Text cloud service out of the services provided by Azure, and AWS. This can be used to help determine what speech recognition service has the best recognition algorithm for different accents as well as speech impediments. The best provider will be determined by accuracy, cost, implementation.

2 OBJECTIVES

The objective will be to find the better of AWS and Azure speech to text cloud service. Each service will get ratings for accuracy, cost, implementation. Each category will be rated then tallied up to produce the final score that will determine the better of the two services. Each category will be rated on the following scale:

- Accuracy: out of 20
- Cost: out of 10
- Implementation: out of 10

Accuracy will be split into two categories: Irish, and Canadian English. The measure we have chosen for accuracy is Word Error Rate (WER). WER is calculated by summing the number of incorrect transcriptions, added words and missing words. That number is divided by the total number of words in the audio clip. With this measure the lower the value the better. A WER of 0 is a perfect transcription. The total score for each category will be $10 - \text{WER}/10$, with a minimum score of 0. Cost will be split into two categories. The first category will be for the free tier the service provider offers. This will be out of five points. This will be one point for every hour of audio than can be transcribed for free each month. The second category will be for the paid tier. It will be out of five points as well. The score for this will be 5 minus value in dollars of one hour of speech transcription. Implementation will be one category with ten points. It will

be a subjective rating based on how easily the service could be used in a python program reading audio in from a file. the rating will be lowered for substantial difficulty implementing the service. Difficulties could be from reading the file in to creating the keys for the service. Difficulties can also be in the form of reading the output. Through this project it was hoped to achieve a substantial result into how well the cloud service works with transcribing speech for any non US accent. This could be the service is very poor at transcribing or excels at transcribing. It was believed before anything in this project was conducted that both services would perform relatively well with US/Canadian accents and poorly with everything but these.

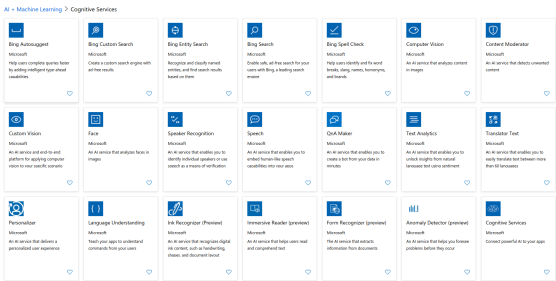
3 CLOUD ASPECTS

3.1 Azure Cognitive Services: Speech

3.1.1 Accuracy

Accuracy is very important when attempting to transcribe speech to text. When using an external service you want to make sure the transcription you are given is accurate. When constantly being given incorrect transcription, why even use the service. The implementation of Azures cognitive services were tested against nine audio clips. Five of these clips were with an Irish accent. Four of these clips were with a Canadian English accent. The output the service produced was directed into a file. That output needed to be cleaned up in order to compare it to the manual transcription. The output for all clips was cleaned from all verbose output. The number of correct, incorrect, added, and missing words were then counted to be able to find the accuracy of each clip. The audio clips with an Irish accent had a WER ranging from 11 percent to 187 percent. This category had both the best and words response. The average WER for audio clips with an Irish accent was 76.15 percent. The audio clips with the Canadian accent had a WER ranging from 30.51 to 66.25. This category had an average of 42.45 percent. The combined average of 61.17 percent. Microsoft stated "A WER of 5%-10% is considered to be good quality and is ready to use. A WER of 20% is acceptable, however you may want to consider additional training. A WER of 30% or more signals poor quality and requires customization and training". (Amos, 2019) There was only one audio clip with a WER less than 30 percent. Based on Azure defined metrics they should have failed. Luckily for them we are not using the same metrics they

Figure 1: List Of Available Services (Microsoft, 2020a)



AI > Machine Learning > Cognitive Services						
Bring Your Own Model Microsoft This lets you bring your own models to the Azure Cognitive Services platform.	Bring Custom Search Microsoft Create custom search engines with Azure Cognitive Services.	Bring Entity Search Microsoft Bring your own entity search models to the Azure Cognitive Services platform.	Bring Search Microsoft Bring your own search models to the Azure Cognitive Services platform.	Bring Spell Check Microsoft Bring your own spell check models to the Azure Cognitive Services platform.	Computer Vision Microsoft An AI service that analyzes content in images.	Content Moderator Microsoft An AI service that detects content that may be inappropriate.
Custom Vision Microsoft An AI service that lets you train custom models to recognize images.	Face Microsoft An AI service that analyzes faces in images.	Speech Recognition Microsoft An AI service that lets you transcribe speech into text.	Speech Microsoft An AI service that lets you synthesize speech from text.	Text Analytics Microsoft An AI service that lets you analyze text for sentiment, topics, and more.	Text Analytics Microsoft An AI service that lets you analyze text for sentiment, topics, and more.	Translator Text Microsoft An AI service that lets you translate text between different languages.
Personalizer Microsoft An AI service that lets you build personalized recommendations.	Language Understanding Microsoft An AI service that lets you understand natural language queries.	Text Analytics (Preview) Microsoft An AI service that lets you analyze text for sentiment, topics, and more.	Image Reader (Preview) Microsoft An AI service that lets you read and understand images.	Form Recognizer (Preview) Microsoft An AI service that lets you extract information from documents.	Anomaly Detector (Preview) Microsoft An AI service that lets you detect anomalies in your data.	Cognitive Services Microsoft An AI service that lets you use a variety of AI services.

are. For Irish accent the average WER of 76.15 gets a score of 2.4. For the Canadian accent the average WER of 42.45 gets a score of 5.8. This gets Azure a total score of 8.2 out of 20. The score for Azure defined acceptable models would be above 16. They are just above half of that.

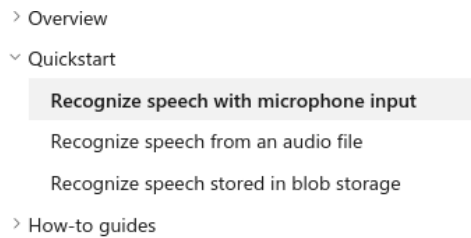
3.1.2 Cost

When performing large projects with many audio clips needing to be transcribed cost can be a big factor as to which service you use. Azure's free tier offers 5 audio hours of transcription each month. For many small projects this is more than enough and can help provide an entry to examining the service. Using these five hours can determine if Azure is the correct service provider for your business. For their free tier they earn all 5 points in the free tier section of the cost category. The paid tier of Azure offers 1 dollar (USD) (Microsoft, 2020b) for each audio hour of transcription. This earns the service 4 points in the paid tier section of the cost service. For cost Azure has managed to grab 9 of 10 points. This is significantly better than the accuracy measure.

3.1.3 Implementation

The implementation of this service was quite pleasant. To start off implementing this service you first need to create the resource in the portal. By searching Speech to Text in the search bar you can either be brought to cognitive services to create the resource or be brought to the speech to text services by Zoom Media based on what you click on. The Zoom media plan is over 100 dollars (USD) a month. This was not the service we wanted. I then had to find Azure's service. I went to Cognitive Services and went through there expansive list. See Figure 1 for the expansive list of offered services. The list was large and finding the right one I wanted took a couple looks. When I first saw speech I thought it only dealt with text to speech, it does however deal with many services related to speech. This is useful if more than one speech service

Figure 2: Tabs Seen For Tutorials (Ryan Hurey, 2020)



is needed however since I only required the one and finding this one service cause some confusion to me Azure lost one point. After creating my resource in Azure I then had to implement the service in Python 3 reading in the audio from files. As a first step I simply googled "Azure Cognitive Services Speech To Text Implementation Python". Clicking the first link brought me to a Azure doc about implementing this service with audio input from my computers microphone. Copying this code into my python file and after downloading the required packages running, I released I forgot to use my key and service region. So going back I added these in and tested out this code from Azure. To my surprise it worked without any errors. Now I just needed to get this to work with files (or so I thought). Going back to this web page i noticed a small tab for reading in from files. See Figure 2 for the tabs. Clicking on this is appeared by using the code given I was done with implementation. I copied the code added one of the audio files in and ran the program. It worked as expected giving me the transcribed text. I then started attempting to transcribe another longer audio clip. After running the program it wasn't transcribing the entire file. I read through the code and there was a comment saying this only works for one statement or 15 seconds. I needed to use the function `start_continuous_recognition()` function instead. I changed the function name and ran the program and it did nothing. Going to Azures docs for the speech sdk proved useless because I couldn't find the function referenced in the docs. That brought me back to google, simply googling the function. I was brought to Azures GitHub samples for speech sdk. By finding the code showing how to use the `start_continuous_recognition()` I used that and ran the program with the longer audio file. It was working but the output had lots of verbose information with it. This output was also being given every time a new word was being recognized. This would have been great if I was attempting to get text as the audio clip was playing. I only wanted the final complete transcription. For this I redirected the output to a file and manually cleaned it. This lost Azure 2 more points. Without looking at how to do it, I am sure i could

have automated this cleaning of the output by using the sdk. The solution to this was not made overly clear which is why I chose to do it manually and why Azure lost the points. If they included cleaning the output to the final complete transcription in the tutorials I was seeing there would not have been point deductions. Overall the implementation took maybe 30 minutes total. About half of this was taken up by myself making a mistake with Python variables and not being able to load a file when trying to automate the redirection of the output and not being required to run the program for each file. For implementation Azure got a 7 out of 10. The implementation was easy to follow and the python packages worked without needing to be played with.

3.1.4 Overall Impression

Azure managed to obtain a score of 8.2 for implementation due to their high WER values. They did however manage to score 9 out of the 10 possible points for cost. For implementation they managed to get a modest score of 7 out of 10. This puts their total score up to 24.2 out of 40. Although I found the implementation process rather easy with great potential for many applications, ultimately they fell very short with accuracy. Not being able to hit the WER they describe as acceptable makes this service fall short. The transcription process was too error prone when making the transcription. If you are using this service to supplement a project to gain some experience using speech to text transcription the free tier is acceptable. Using this in a service that requires accurate transcriptions would be a different story. The transcriptions would not perform well enough to be able to be used. The cost of using the service would grow trying to attempt to get an accurate transcription. The time required to do some of the cleaning on the response would incur costs. Overall the impression of Azure Cognitive Services: Speech was that the service still has far to go before it should be used in production environments.

3.2 AWS

3.2.1 Accuracy

Amazon Transcribe finished with an average accuracy of 60.1 percent, across nine audio files of strong Irish and Canadian accents. This was captured by calculating the average word error rate per program and averaging them out across all tests. The word error rate was the total amount of missed words, added words and wrongly translated words, over the total amount of words in the audio file. The service scored

a WER of 58.3 (58.3 percent of the transcription was wrong) with the Irish accent clips. While it scored WER of 21.5 when tested with the Canadian accents. These score allowed us to get a greater understanding of the effectiveness of the speech to text tool that AWS offers. The accuracy of these tests depended on lots of variables. These variables were partly effected by us, the experimenters, and partly by the service itself. An example of a factor that effected accuracy that we controlled was the manual transcriptions that we were comparing against. Seeing as these transcriptions were done by us humans, and the difficulty of understanding the audio clips we were using, it was impossible to confidently say that our manual transcriptions were 100 percent accurate. This creates a flaw in the experiment and could lead to inaccurate accuracy numbers.

Additionally, Amazon transcribe offered language support for different accents across many languages. This could be a large factor that accounts for the out-performance by nearly 20 percent of Azure. Some languages/accents it currently supports are (Kranz, 2020):

- Gulf Arabic (ar-AE)
- Chinese Mandarin - Mainland (zh-CN)
- Irish English (en-IE)
- US English (en-US)
- Welsh English (en-WL)

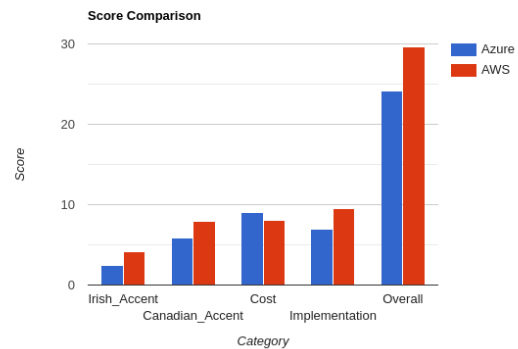
3.2.2 Cost

Similar to other AWS Services, Amazon Transcribe offers a free-tier pricing model where they offer a free 1 hour of services per month for the first year. The advantage of AWS offering this free-tier is that it gives users the opportunity to test out the service and compare it to others before implementing it in their software. The cost of 1.44 per hour of transcribing is 44 percent higher than Azure and this would be extremely significant when you require hundreds or thousands of hours of transcribing.

3.2.3 Implementation

The implementation of Amazon Transcribe was seemingly effortless. AWS offers a SDK for Python called boto3 that is the same SDK used for many of it's other services as well. This simplifies the implementation process and allows for the combination of multiple AWS services into one. In our implementation of Amazon Transcribe, we used boto3 to access

Figure 3: Score Comparison For Azure and AWS



S3 which is a storage within AWS that held the audio clips, integrated with Transcribe. On top of the SDK, Amazon Transcribe offers an extremely straight forward interface that allows users to create jobs but with the click of a few buttons and explore the many growing features this service offers. Users are able to test out the real-time implementation directly on the AWS Console, as well as create custom vocabulary and filtering for specified vocabulary as well.

3.2.4 Overall Impression

Overall the best part of Amazon Transcribe was the implementation. Similarly to other AWS Services, Transcribe provides a great experience and an easy to use service. The worst aspect of Amazon Transcribe was it's accuracy rating against complex audio files. With an average of 40 percent error this was simply unacceptable for the transcription of any speech. We are confident that with time and more machine learning, this number will improve. Lastly, the cost for this service is quite high compared to it's Azure counterpart, but this could be the cost worth paying for a more accurate service with a better implementation.

4 MAJOR ACCOMPLISHMENTS AND CONTRIBUTIONS

4.1 Accomplishments

During the project there were a few important discoveries made. With respect to Azure it was discovered that the 20 percent acceptable WER for a model was unreachable for the standard model deployed. This is significant as you would expect a service provider to be able to reach the scores of what they have described as a deploy-able model. In the comparison of

the two services AWS had the better score of 4.2 to Azure's 2.4 for the Irish accent category. Although it is 1.8 points higher, AWS managed to nearly double Azure's score. This shows that AWS is better than Azure but, neither service is rather well at transcribing audio clips with Irish accents. AWS managed to best Azure again in the Canadian accent category with a score of 7.9 to Azure's 5.8. This is again about 2 points higher. While Azure's score indicates that it needs improvement for transcriptions, AWS has a score to what can be a passable model. The score is high enough that given a non mission critical scenario the model could be used to transcribe audio to text. Azure won its only category in cost by being slightly cheaper to transcribe text per audio hour. The win was by 1 point, a score of 9 to AWS' 8. Cost is an important factor however you may not want to pay for a service that is performing sub-par. AWS Pulled farther ahead of Azure in implementation with a score of 9.5 to 7. Azure's slight difficulties with creating the service and managing output was slightly harder than AWS. Overall AWS won with a score of 29.6 to 24.2. There was a 5.2 point difference between the services however with a poor performance in accuracy Azure simply could not keep up with AWS. See figure 3 for a chart comparing the scores.

4.2 Contributions

For this project we had to learn how to compare the accuracy of transcribed audio files. Using resources found on Azure's docs we found a comparison measure that helps to accurately compare the number of errors made. We each needed to learn how to use and implement one of Azure or AWS speech to text transcription service.

contributions can be small items that you might not have planned for but had to do or learn to aid in your project work.

5 FUTURE WORK

We believe our approach to determining which service was the "best" was rather sound. There are some things that we could definitely improve upon that would improve the results and provide a stronger foundation for our findings. Given more time we would have liked to include more cloud services such as Google, Alibaba, and IBM. This would provide more choice for seeing where different services are better than others. By only comparing AWS and Azure another company that wasn't looked at may have a more complete model that better encompasses

different types of speech. With more time, a larger data set for each of the speech types we have would be able to provide more reliable statistics on accuracy. By chance we could have either picked words that the models excel at or fail at. This can misrepresent the true accuracy of the services compared. A large improvement that could be made is better transcriptions. With the choice of Irish accents it was difficult to get the correct transcription of what was being said in the audio clips. This can effect the calculated accuracy of the models. If the model got the word correct but the manual transcription is wrong the model will have a lower accuracy than it should. If the model gets the word wrong but matches the incorrect transcription the accuracy will be incorrectly increased. Lastly, more types of speech could have been used. We found it difficult to find audio clips of natural speech where people were not attempting to mimic the types of speech we were looking for. We didn't want to use these audio clips as mimicking the type of speech could improperly represent the model with small inconsistencies.

6 TEAM MEMBER CONTRIBUTIONS

6.1 Team Member: Tyler Green

During this project I had many roles. I was in charge of obtaining audio clips. These clips had to be long enough that there was more than one or two words. These clips also had to have easily heard speech. We didn't want to try and trick the models by more than just using different speech types. Once the audio clips were found I then had to transform them into the common wav format. This was because I was obtaining the clips in mp3 format. Although Azure does support mp3 format it was easier to convert the audio clips and use them in wav format. I was then in charge of manually transcribing the audio clips. This was a long and difficult task for the clips with Irish accent. Some of the audio clips had easier to understand accents making those easy, however the heavier accents were a difficult task to transcribe. There was one clip where a full transcription was not possible and only part of the clip could be used. After the clips were found and transcribed my task was to perform the Azure implementation. After being done the implementation and using the implementation to obtain Azure's transcriptions of the audio files. It was my task to clean the output. After having cleaned output I created the metrics used to provide accuracy and a score rating of each

attribute we were evaluating. I then started working on the report and slides for the project. I completed the introduction, objectives, Azure tutorial, major accomplishments and contributions, and future work of the document. I completed slides 2, 3, 7, 10, 12. I completed azure demo.

6.2 Team Member: Loui Zibdawi

Throughout the planning and execution of this assignment I had the role of researching and testing Amazon Transcribe. This involved setting up the python script that used boto3 to transcribe all the audio files in a specified S3 bucket. I stayed in communication with my partner, Tyler, through Slack and often touched based throughout milestones of the assignment. I was responsible with putting together all the audio recordings with the slides as well as creating the Amazon Transcribe demos. Additionally, I completed the slides for contributions of AWS, the cloud aspects of AWS, major accomplishments of AWS and what we would change if we did it all again.

REFERENCES

- Amos, M. (2019). What is word error rate (wer)? In <https://docs.microsoft.com/bs-cyrl-ba/azure/cognitive-services/speech-service/how-to-custom-speech-evaluate-data>. Microsoft Azure.
- Kranz, G. H. (2020). What is amazon transcribe? In <https://docs.aws.amazon.com/transcribe/latest/dg/what-is-transcribe.html>. Amazon Transcribe.
- Microsoft (2020a). Cognitive services. In https://portal.azure.com/#blade/Microsoft_Azure_Marketplace/MarketplaceOffersBlade/selectedMenuItemId/CognitiveServices_MP/dontDis. Microsoft Azure.
- Microsoft (2020b). Cognitive services pricing—speech services. In <https://azure.microsoft.com/en-us/pricing/details/cognitive-services/speech-services/>. Microsoft Azure.
- Ryan Hurey, Rob Chambers, D. P. e. (2020). Quick-start: Recognize speech from a microphone. In <https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/quickstarts/speech-to-text-from-file?tabs=linux&pivots=programming-language-csharp>. Microsoft Azure.