Module 8: It's the Final Countown!

IFT 598: Natural Language Processing

Speech Recognition Project

Jaime Baeza

Chandana Kandari

Rohitha Jakkidi

4/23/2021

**Table of Contents**

**Abstract...........................................................................................................................................4**

Briefly define the question that you are asking ………………………………………………4

**Question..........................................................................................................................................5**

What is the question you are answering? ..................................................................................5

Why is it important?..................................................................................................................5

Articulate the context of the question........................................................................................5

How are you going to assess the process and outcome of your project? ..................................6

**Data Acquisition.............................................................................................................................7**

Identify the data source..............................................................................................................7

Describe the iterative nature of NLP in relation to your question and deliverable. ..................8

**Data Cleaning.................................................................................................................................9**

Clearly identify the process that you would use to manage the data.........................................9

Identify the process and methodology you would use to "clean" the data................................9

**Data Modeling..............................................................................................................................10**

Identify the type of model that would prove most advantageous to your question and deliverable….10

**Data Analysis................................................................................................................................11**

What should the product look like at the analysis phase?........................................................11

How would you measure your success?..................................................................................16

What "unexpected outcomes" might you anticipate ...............................................................16

**Conclusion....................................................................................................................................16**

What conclusions can be made? .............................................................................................16

How can this be integrated?.....................................................................................................18

What is the value added? ........................................................................................................18 **Appendices....................................................................................................................................19**

Appendix 1: .............................................................................................................................19

Appendix 2: .............................................................................................................................21

Appendix 3: .............................................................................................................................22

Appendix 4: .............................................................................................................................25

**References.....................................................................................................................................28**

**Abstract**

**Briefly define the question that you are asking**

Speech recognition is a vital tool that allows for the development of technology with the capacity of recognizing human speech that enables devices to adapt voice information in a form that it can understand. Speech recognition systems will ease communication between technology and humans, in its more basic form “has a limited vocabulary of words and phrases and may only identify these if they are spoken very clearly. More sophisticated software has the ability to accept natural speech” (Husnjak, Perakovic, & Jovovic, 2014).

The significance of this project is the added knowledge that will come from the vigorous testing done by analyzing the correctness of textual output obtained from Google Speech recognition API to determine if the API can accurately determine words spoken from people on various criteria such as birthplace of speaker and native language. The purpose is to evaluate the accuracy of speech recognition tools at identifying human language spoken in order to help improve the accuracy of the speech recognition tool and make it useful and reliable for everyone. Using the results of our analysis of the common trends in misidentifying words pronounced in different actions we can help improve voice recognition by improving and noting noticeable mistakes and can implement solutions to improve distinguishing the language.

As a final product we are going to present a dashboard with visualizations that shows how effective the Speech recognition tool can identify the words under given conditions like variability in accent, connotations, vocal frequency ranges present in input audio files depending on various criteria such as age, birthplace of speaker, native language, and sex. This analysis provides insights for comparing the demographic and linguistic backgrounds of the speakers in order to show that accents can be identified.

**Question**

**What is the question you are answering?**

The question that will be answered is: How effective is NLP in recognizing speech, and can it perform a task like identify or translate speech to text? This Project will determine how effective Speech recognition tools like the Google API are in identifying words spoken by people with different accents under given conditions present in input audio files depending on various criteria. This will be done for the purpose of evaluating the accuracy of speech recognition tools at identifying human language spoken.

**Why is it important?**

This question is important because it recognizes the importance that speech recognition is in society. Speech recognition has “widespread application in tasks that require a human-machine interface, such as automatic call processing in the telephone network and query-based information systems that do things like provide updated travel information, stock price quotations, weather reports, etc.” (Juang & Rabiner, 2005). The significance in analyzing the effectiveness of textual output obtained from Google Speech recognition API will help determine if the tool can effectively determine what a person says.

**Articulate the context of the question.**

Stakeholders are individuals such as people working with NLP, machine learning programmers, text to speech programming, etc. To involve stakeholders a report of findings in NLP applications in speech recognition can be given to give them a better understanding of its importance. Natural language processing is one of the emerging technologies that has become a major part of current research in linguistic analysis and processing. It is “an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things” (Chowdhury, 2003).

One of the goals that NLP wants to accomplish is to understand how humans understand, use, and interact with language for the purpose of developing computer systems that can aggregate the different components of language in order for it to understand it. “Applications of NLP include a number of fields of study, such as machine translation, natural language text processing and summarization, user interfaces … speech recognition, artificial intelligence, and expert systems” (Chowdhury, 2003). The ultimate goal is to get computers to the point that they are able to achieve human-like comprehension of texts or languages. The variables that are used in the project will be the Test audios, output files, speech text analysis methods, formatting, variability in accents, etc.

NLP is the best tool to respond to this question. Natural language processing is so integral due to it using text analysis, computational linguistics, and structures of language that allow for speech recognition to be implemented. Using various sources of datasets such as Kaggle can be used as a research tool to perform the project and by using their expansive sample files to test the functionality of the project. Other data supports the question as there has been different research in speech recognition that has data available to use.

**How are you going to assess the process and outcome of your project?**

To assess the process, tests will be used to see if speech recognition is performed to identify and translate incoming inputs with the outputs being able to be seen to assess how effective the project is. As a final product we are going to present a dashboard with visualizations that shows how effective the Speech recognition tool is. This analysis provides insights for comparing linguistic backgrounds of the speakers and will show if accents can be identified by the program or if it has a difficult time doing so.

**Data Acquisition**

**Identify the data source.**

The project will make use of tools that analyze speech and convert it into text using audio files as an input. The audio files will be from a dataset that was collected by Steven H. Weinberger being hosted by George Mason University. (Weinberger, 2013). This dataset contains 2140 speech samples, each from a different talker reading the same reading passage. Talkers come from 177 countries and have 214 different native languages with each talker speaking a sentence of text in English. The project will collect data on the effectiveness that our implementation of speech recognition using python has at accurately translating spoken words and converting it into text. This data will be collected ourselves and data that we will collect will show how effective the tool is at identifying the words

There are several APIs available to convert text to speech in python. We will be using Google Speech recognition API. When performing Speech Recognition of an audio source from a file, we will give a path to our Audio File we want to transcribe to text as an input to google speech and then it will give us the textual output which will print out to the Screen. We have to use the Recognizer class available in the speech\_recognition module to translate the speech into text.

The audio file path will be passed to the AudioFile class constructor. We will use the recognize\_google () method for the transcription of the audio files. However, the recognize\_google () method needs a parameter. We can use the record () method available in the Recognizer class in order to translate the audio file to an audio data object. We have also used the recognize\_google () method for using the Google API for translation, and we can observe the output that the audio has been successfully converted into text. In our project, we are going to analyse the reasons for the inaccuracy caused by certain data that is not transcribed using an accuracy method in the code to compare the text to the original sentence that was supposed to be spoken.

**Describe the iterative nature of NLP in relation to your question and deliverable.**

Natural language processing has an iterative nature that is used in the project. The program iteratively finds the audio files from a folder and passes each of them through the google API which returns the results of its analysis of the words spoken and translates it into text. This is then passed back into the program where an output file is developed that shows the audio file name and the text result. From there we used another NLP iterative method which involves comparing the text to the original sentence and getting a percentage of the accuracy that the google API was able to translate the audio file.

Through the code this process is done iteratively for all of the audio files found on the google drive and the results are neatly outputted in an xml or excel file that can then be used for data visualizations. By the nature of this project an iterative approach was needed due to the high volume of audio test files used in the project to test the Google speech recognition and see how accurate it is when compared to the different accents of the speakers. The strengths are it uses already existing speech to text recognition that reads in the files iteratively and gets back a result in the same way.

**Data Cleaning**

**Clearly identify the process that you would use to manage the data.**

Data will be collected by testing different audio files from the database to see our tools effectiveness and will collect our findings so that we can create data visualizations that will help show off the project's results. To collect the data, we will use an output file produced by the code in an excel file. To get accurate results we would need to perform a large quantity of samples of different variables to get a wide diversity in demographics so that our data is accurate. The database is quite large and it fairly and accurately represents the dataset as a whole. The dataset that will be used in this project contains some sensitive data due to it having demographic data on people who participated in the project such as their name, sex, gender, etc. In order to protect individuals in this dataset from references to anyone it will be through the use of a speaker ID which is given in this dataset so that people won’t know who is being talked about.

**Identify the process and methodology you would use to "clean" the data.**

In order to clean the data that we gather we will look at each of the rows in the output excel file and determine which points of data are not helpful in our analysis and we can remove them. Some of the files were not able to be processed through the program so in the analysis these entries will be removed. Once the data that we collected from converting the database of audio files to text we are able to create the data visualizations that show the accuracy that the google API had for transcribing the sentence spoken and compare it to the original sentence to see the percent accuracy of the API. We can then do analysis based on various parameters.

**Data Modeling**

**Identify the type of model that would prove most advantageous to your question and deliverable.**

In order to do data analysis and visualizations with the output file that has the accuracy information on the Google API and how effectively it was able to identify speech of people with accents and convert it into text we need to use a data modeling tool. The tool used for our analysis is Tableau which is a data visualization software that provides features such as query relational databases, cloud databases, and spreadsheets to generate graph-type data visualizations. Using Tableau, we can create a dashboard of different comparisons between a demographic and the average percentage accuracy of the Google API in speech recognition. Using Tableau, we cave choose to do bar charts as a way of modeling the data that can change depending on the parameters, we placed such as age groups, sex, and filter by country.

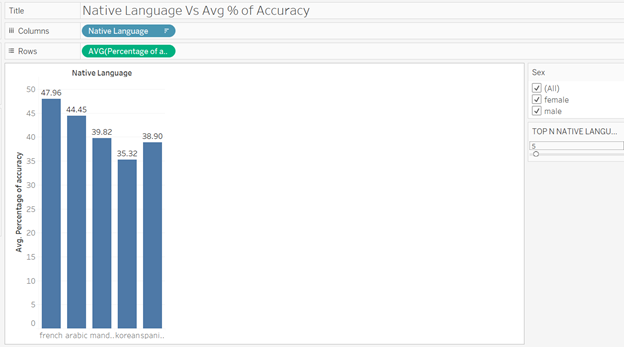
This model is optimal for this situation as it makes displaying the comparison between the accuracy easier to visualize and compare with each other. Some of the strengths and weaknesses to this is that it is limited to generalizations as it's not optimal to display individual scores of people and can only do a large comparison. For this project we created a file for the comparisons that uses python code to perform natural language processing by using the google speech recognition API. Python was a great tool to use as it is a very robust language that can do all sorts of functions from machine learning to manipulating data for various purposes.

**Data Analysis**

**What should the product look like at the analysis phase?**

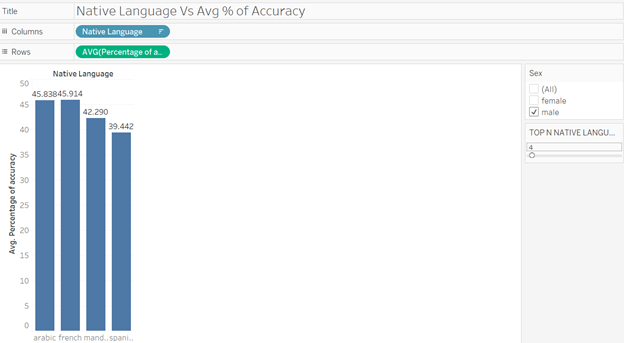
**Visualization – 1**

We have visualized the Average percentage of accuracy over the Native language of the speaker. Here, in this visualization we have used a Bar Chart to represent the average percentage of accuracy of the speaker based on their native language. We have also provided filters for selecting TOP N native languages that have the highest average percentage of accuracy of the speech to text analysis.

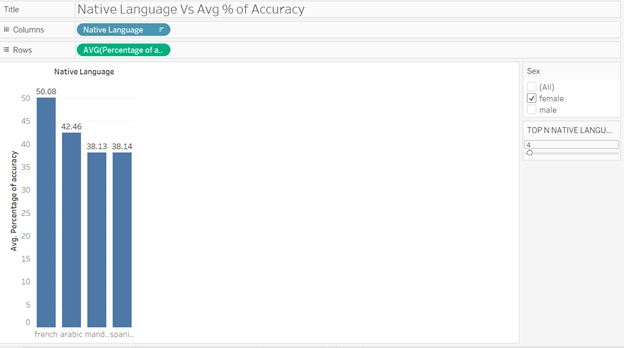


We can also analyze the average percentage of accuracy of the speaker based on the gender of the speakers so that we could compare the analysis of the accuracy provided by male and female speakers.

1. As you can see, if a male speaker is selected the top 4 Native languages that hold the highest average percentage of accuracy of the speech to text analysis are Arabic, French, Mandarin, Spanish.



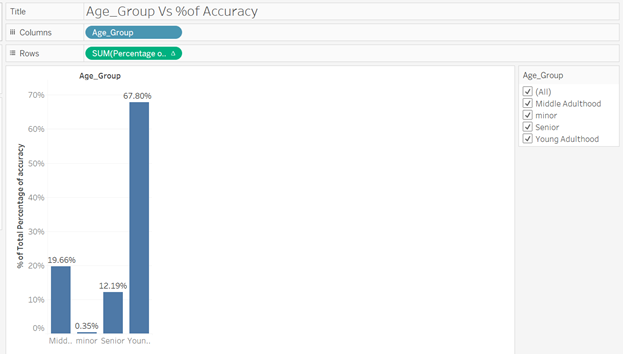
2. As you can see, if a female speaker is selected the top 4 Native languages that hold the highest average percentage of accuracy of the speech to text analysis are French, Arabic, Mandarin, Spanish.



**Visualization – 2**

We have visualized the Average percentage of accuracy over the Age Group that the speaker falls in. Here, in this visualization we have used a Bar Chart to represent the average percentage of accuracy of the speaker based on their Age Group. We have created the calculated field as below to classify the Age Group.



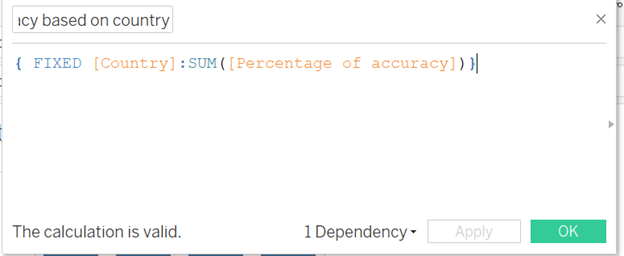


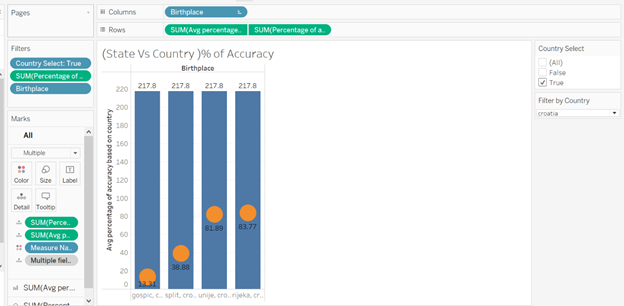
This Visualization helped us to understand the total percentages of accuracy the speakers of certain age groups hold. As we can see in the results, the Young adulthood holds the high precision of accuracy of the speech to text analysis under given conditions like variability in accent, connotations, vocal frequency ranges present.

**Visualization – 3**

We have visualized the Average percentage of accuracy of the speakers belonging to different regions in a particular country. How much is the region or state contributing to the accuracy in comparison to the country’s total percentages of accuracy.

We have provided the Filter for selection of a country and a calculated field for measuring the Country’s average percentage of accuracy. Which is plotted against the percentages of accuracy provided field obtained after performing the analysis of accuracy on speech to text output.





In the above visualization, we can see the contributions for the accuracy by speakers of each region of Croatia with respect to their country Croatia. So, the speakers belonging to the Gospic region possess low accuracy in their delivery of speech compared to speakers of other regions of Croatia.

**How would you measure your success?**

We say that we were to successfully perform analysis on how effective NLP is in recognizing speech, and can it perform a task like identify or translate speech to text? This Project was able to use the Google API in identifying words spoken by people with different accents for the purpose of evaluating the accuracy of speech recognition tools at identifying human language spoken. We measured our success by developing a dashboard of different comparisons based on different parameters such as age, country, and sex. We are able to make different comparisons and determine how successful the API was able to identify accents.

**What "unexpected outcomes" might you anticipate**

Some of the unexpected outcomes that we anticipated were some of the percentages in the accuracy of the conversion where there will be some unexpected results. We also anticipated the conversion to be somewhat unreliable as the results varied each time we compiled so after a few trials once we saw similar results we stopped and produced the output file that was then used to do the data analysis.

**Conclusion**

**What conclusions can be made?**

We have visualized the Average percentage of accuracy over the Native language of the speaker. As seen in the analysis for male speakers the top 4 average percentages of accuracy of the speech to text analysis are Arabic, French, Mandarin, Spanish. While for females it's the same languages French, Arabic, Mandarin, Spanish. This means that the API has the easiest time picking up accents of both males and females who speak those four. The API from what we saw is able to analyze speech from both males and females with the same or similar accuracy. For the Average percentage of accuracy over the Age Group with the results showing that Young adulthood holds the high precision of accuracy of the speech to text analysis. This means that the API is better at recognizing words spoken by specific age groups while anyone over or under has more difficulty in deciphering the speech.

Lastly, the average percentage of accuracy of the speakers belonging to different regions in a particular country. We saw that speakers belonging to the Gospic region possess low accuracy in their delivery of speech compared to speakers of other regions of Croatia.This means that even though the people are from the same countries their accents are different between regions which produces different accents where the API can understand some accents better then others. This Trend is shown through all of the other various countries where regions make a difference in people's accents so the API may have a difficult time picking up all of these variations.

Due to our results it shows that the biggest difference in translating speech has to do with the many accents produced from regional differences between the same group of individuals living in the same country. It also showed preference to younger age groups when it tries to understand the words spoken. Due to these results some outcomes that can come from making this analysis is branching out more to some of the variation found in different regions as they are important in trying to improve speech recognition systems.

**How can this be integrated?**

The significance of this project is the added knowledge that came from our testing and how we analyzed the Google Speech recognition API to determine if the API can accurately determine words spoken from people with accents. Using the results of our analysis of the common trends of the API, misidentifying words pronounced in different accents can be integrated as a sort of usability testing that can help enhance the Google speech recognition API. We can help improve voice recognition by improving and noting noticeable mistakes and can implement solutions to improve distinguishing the language. The main thing we found is the importance of getting the linguistic variations of accents due to regions where people live as this subtle difference can show different scores of accuracy in translating from speech to text.

**What is the value add?**

To assess the process, analysis was used to identify the accuracy of translating incoming inputs with the outputs being used to test this. This analysis provides insights for comparing linguistic backgrounds of the speakers and will show if accents can be identified by the program or if it has a difficult time doing so. This is the value added from NLP, which enhances the accuracy of speech recognition tools at identifying human language and will help improve the accuracy of the tool. Due to these results branching out more to some of the variations found in different regions is important in trying to improve speech recognition systems.

**Appendices**

**Appendix 1: Code for the audio to text conversion**

import openpyxl

import speech\_recognition as sr

from os import path

wb=openpyxl.load\_workbook('speakers\_all.xlsx')

sheet=wb.get\_sheet\_by\_name('speakers\_all')

n=sheet.max\_row

for i in range (2,n+1):

print (i)

# print(sheet.cell(row=i,column=4).value)

s=sheet.cell(row=i,column=4).value

sound=s+".wav"

try:

audio\_file = path.join(path.dirname(path.realpath(\_\_file\_\_)), sound)

r=sr.Recognizer()

with sr.AudioFile(audio\_file) as source:

audio = r.record(source)

try:

#print (r.recognize\_google(audio))

sheet['j{}'.format(i)].value=r.recognize\_google(audio)

except sr.UnknownValueError:

# print("can't understand")

sheet['j{}'.format(i)].value="not recognised"

continue

except:

sheet['j{}'.format(i)].value='No audio file exists with that name'

continue

wb.save('testing2.xlsx')

**Appendix 2: Code for the percent accuracy**

import openpyxl

import difflib

wb=openpyxl.load\_workbook('testing2.xlsx')

sheet=wb.get\_sheet\_by\_name('speakers\_all')

n=sheet.max\_row

source="Please call Stella ask her to bring these things with her from the store six spoons of fresh snow peas five thick slabs of blue cheese and maybe a snack for her brother Bob we also need a small plastic snake and a big toy frog for the kids she can scoop these things into three red bags and we will go meet her Wednesday at the train station"

for i in range (2,n+1):

temp=sheet['j{}'.format(i)].value

print(temp)

result=difflib.SequenceMatcher(None,source,temp).ratio()

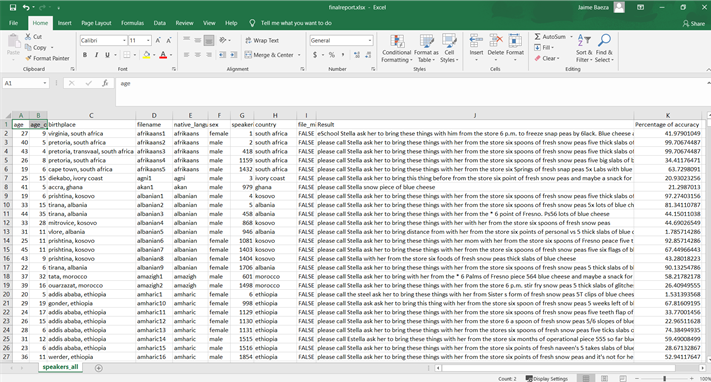
sheet['k{}'.format(i)].value=float(result)\*100

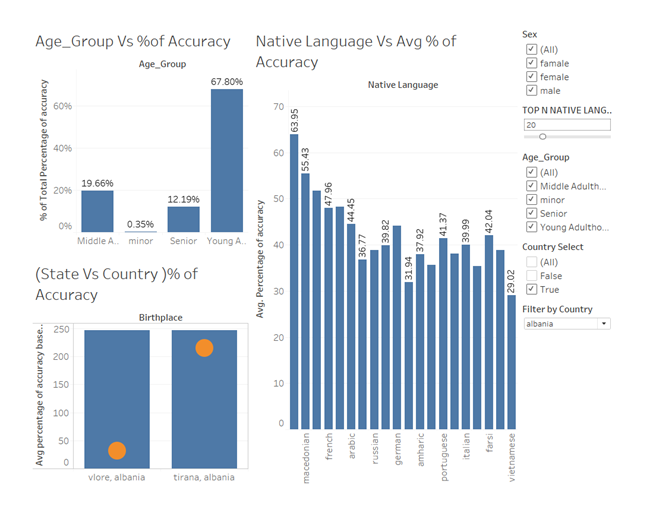
# print(float(result)\*100)

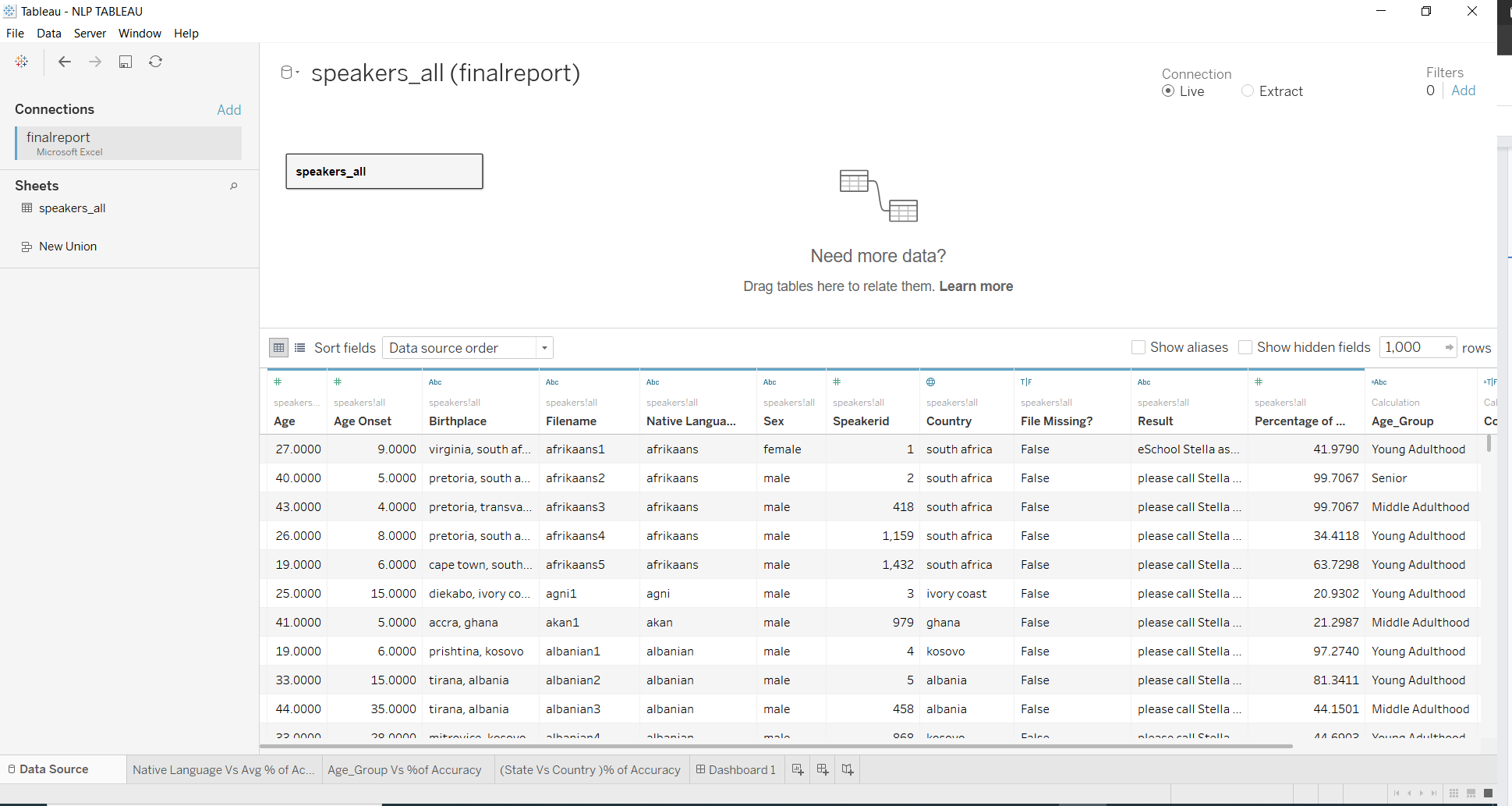
wb.save('finalreport.xlsx')

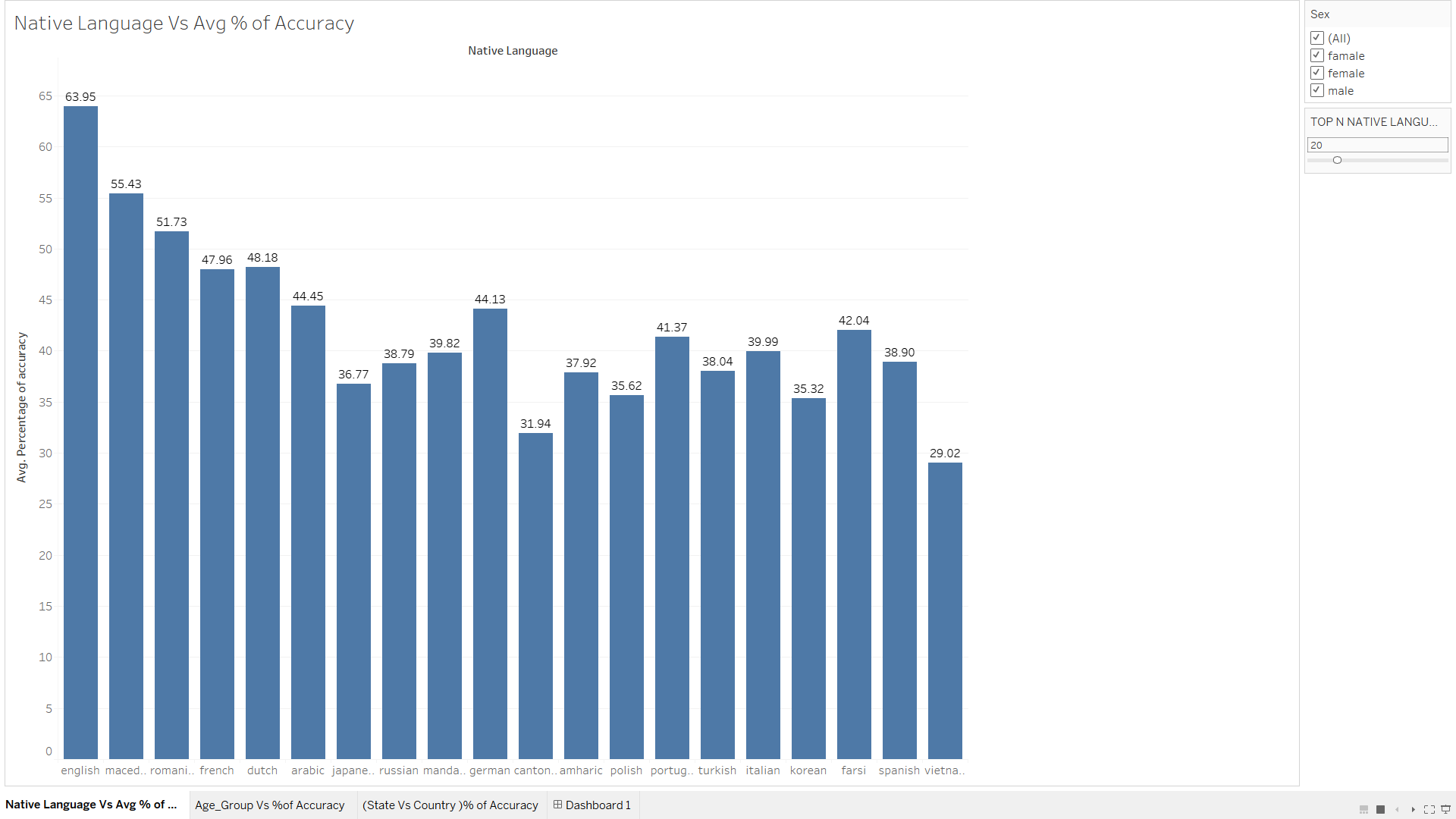
**Appendix 3: Sample output file with audio files database information, result of audio to text, and accuracy comparison to original sentence.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| age | birthplace | filename | native\_language | sex | country | Result | Percentage of accuracy |
| 27 | virginia, south africa | afrikaans1 | afrikaans | female | south africa | eSchool Stella ask her to bring these things with him from the store 6 p.m. to freeze snap peas by 6lack. Blue cheese and maybe a snack for her brother brother but we also need a small plastic snake and a big boy frog for the kids she can screw these things into free weight back and we will go meet Wednesday at train station | 41.97901049 |
| 40 | pretoria, south africa | afrikaans2 | afrikaans | male | south africa | please call Stella ask her to bring these things with her from the store six spoons of fresh snow peas five thick slabs of blue cheese and maybe a snack for her brother Bob we also need a small plastic snake and a big toy frog for the kids she can scoop these things into three red bags and we will go meet her Wednesday at the train station | 99.70674487 |
| 43 | pretoria, transvaal, south africa | afrikaans3 | afrikaans | male | south africa | please call Stella ask her to bring these things with her from the store six spoons of fresh snow peas five thick slabs of blue cheese and maybe a snack for her brother Bob we also need a small plastic snake and a big toy frog for the kids she can scoop these things into three red bags and we will go meet her Wednesday at the train station | 99.70674487 |
| 26 | pretoria, south africa | afrikaans4 | afrikaans | male | south africa | please call Stella ask her to bring these things with her from the store six spoons of fresh snow peas five big slabs of blue cheese and maybe a snack for her brother Bob we also need a small plastic snake and a big toy frog for the kids she can scoop these things into three red bags and we will go meet her Wednesday at the train station | 34.41176471 |
| 19 | cape town, south africa | afrikaans5 | afrikaans | male | south africa | please call Stella ask her to bring these things with her from the store six Springs of fresh snap peas 5x Labs with blue cheese and maybe a snack for her brother Bob we also need a small plastic snake in the big toy frog to the kids she can skip these things into three red bags and we were going to meet her Wednesday at the train station | 63.7298091 |



**Appendix 4: Data Visualizations Dashboard** 









**References**

Chowdhury, G.G. (2003), Natural language processing. Ann. Rev. Info. Sci. Tech., 37: 51-89. <https://doi.org/10.1002/aris.1440370103>

Husnjak, S., Perakovic, D., & Jovovic, I. (2014, March 25). Possibilities of using speech recognition systems of smart terminal devices in traffic environment. Retrieved March 27, 2021, from<https://www.sciencedirect.com/science/article/pii/S1877705814003002>

Juang, B. & Rabiner, Lawrence. (2005). Automatic Speech Recognition - A Brief History of the Technology Development. [Automatic Speech Recognition - A Brief History of the Technology Development | Request PDF (researchgate.net)](https://www.researchgate.net/publication/249888949_Automatic_Speech_Recognition_-_A_Brief_History_of_the_Technology_Development)

Weinberger, S. (2013). Speech accent archive. George Mason University. [Speech Accent Archive | Kaggle](https://www.kaggle.com/rtatman/speech-accent-archive)

**Code helper links:**

* <https://pypi.org/project/openpyxl/>
* <https://stackoverflow.com/questions/52283840/i-cant-install-pyaudio-on-windows-how-to-solve-error-microsoft-visual-c-14>
* <https://stackoverflow.com/questions/10849141/can-i-do-a-string-contains-x-with-a-percentage-accuracy-in-python>
* <https://pypi.org/project/SpeechRecognition/>
* <https://github.com/Uberi/speech_recognition/blob/master/examples/audio_transcribe.py>