B20AI013_SU_PA1_REPORT

By: **Ishaan Shrivastava [B20AI013]**

> 💡 Note: In this lab assignment, I will attempt to **gain an understanding** of the various MMS LID and ASR models mentioned, as well as **documentation of any interesting or notable findings (explanation of output, etc.).** Any feedback on my work or how to improve my submission is highly appreciated.

Question 1: Introduction to Audio Applications

Task 0: Recording two sentences in English and your native language in your own voice.

For the purpose of testing the performance of the MMS-LID, MMS-TTS, and MMS-ASR models, I have recited myself speak two paragraphs, each in:

1. **Hindi:** Native language. Original language of source paragraphs.
2. **English:** Paragraphs originally in written in Hindi, translated to English and recorded.

Recording	Length
1hin.wav	0:19
2hin.wav	0:19
1eng.wav	0:16
2eng.wav	0:14

Note: I have already converted some of the recordings to the desired sampling rate and .wav format on my local device. These files are also available inside the zip file submitted with the report.

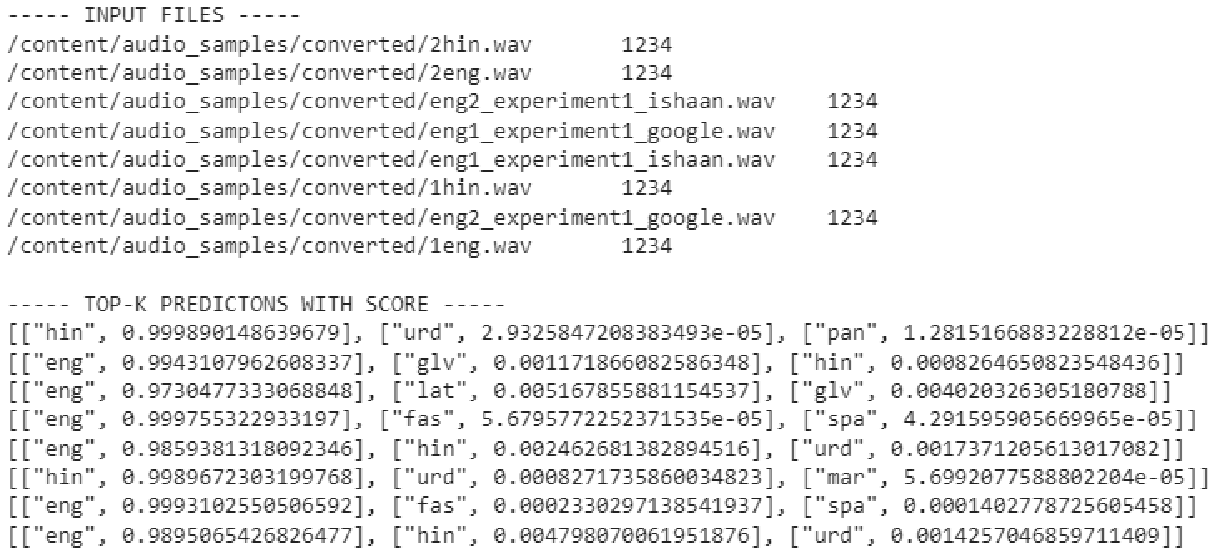
Task 1: MMS-LID models - Accuracy (predicted vs ground truth languages, performance analysis)

MMS-LID stands for Massively Multilingual Speech Language Identification. It is a series of models by Facebook, that do multilingual speech identification by mapping each audio input to a probability distribution over multiple output classes.

There are trained MMS-LID models that classify across 126, 256, 512, 1024, 2048, and 4017 output languages respectively.

MMS-LID models use the wav2vec 2.0 architecture, pretrained using a contrastive loss task and then finetuned for language identification.

Using the colab files given, I ran the speech identification task on the above recordings. Here are the results: (confidence score is with respect to the correct language for the given recording, hindi in case of 1hin.wav and 2hin.wav, and English in case of 1eng.wav and 2eng.wav)



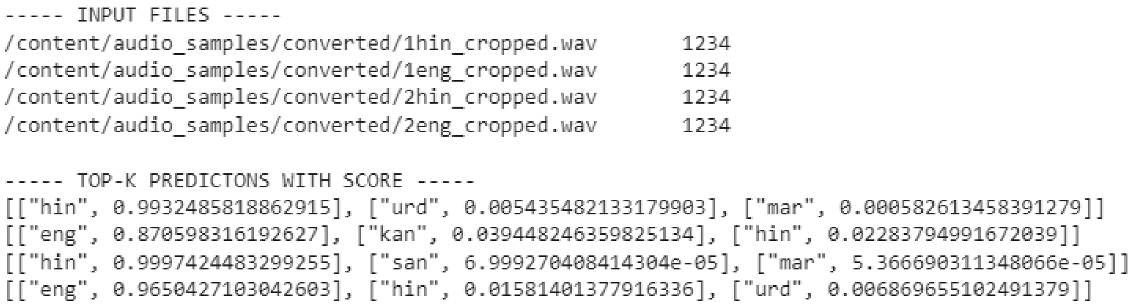
```
----- INPUT FILES -----
/content/audio_samples/converted/2hin.wav       1234
/content/audio_samples/converted/2eng.wav       1234
/content/audio_samples/converted/eng2_experiment1_ishaan.wav    1234
/content/audio_samples/converted/eng1_experiment1_google.wav    1234
/content/audio_samples/converted/eng1_experiment1_ishaan.wav    1234
/content/audio_samples/converted/1hin.wav       1234
/content/audio_samples/converted/eng2_experiment1_google.wav    1234
/content/audio_samples/converted/1eng.wav       1234

----- TOP-K PREDICTONS WITH SCORE -----
[["hin", 0.999890148639679], ["urd", 2.9325847208383493e-05], ["pan", 1.2815166883228812e-05]]
[["eng", 0.9943107962608337], ["glv", 0.001171866082586348], ["hin", 0.0008264650823548436]]
[["eng", 0.9730477333068848], ["lat", 0.005167855881154537], ["glv", 0.004020326305180788]]
[["eng", 0.999755322933197], ["fas", 5.6795772252371535e-05], ["spa", 4.291595905669965e-05]]
[["eng", 0.9859381318092346], ["hin", 0.002462681382894516], ["urd", 0.0017371205613017082]]
[["hin", 0.9989672303199768], ["urd", 0.0008271735860034823], ["mar", 5.6992077588802204e-05]]
[["eng", 0.9993102550506592], ["fas", 0.00023302971385419537], ["spa", 0.0001402778725605458]]
[["eng", 0.989506542682647 7], ["hin", 0.004798070061951876], ["urd", 0.0014257046859711409]]
```

Recording	Length	Confidence score
1hin.wav	0:19	0.9989
2hin.wav	0:19	0.9998
1eng.wav	0:16	0.9895
2eng.wav	0:14	0.9943

The MMS model predicts the correct language with a high confidence for all four samples. It can also be noted that the model gives higher confidence scores for hindi recordings compared to english ones, as the error rate is atleast an order of magnitude lower. I suspect that this could be due to my accent while speaking english not being a native one, which could lower the confidence of the model.

Experiment 1: confidence score variation with recording length

To figure out why the confidence scores were so high even for a model with so many classes, I tried to reduce the length of the audio recordings by cropping them to 6 seconds each. Here is the list of audio recordings with their lengths and confidence scores, as well as the originals.



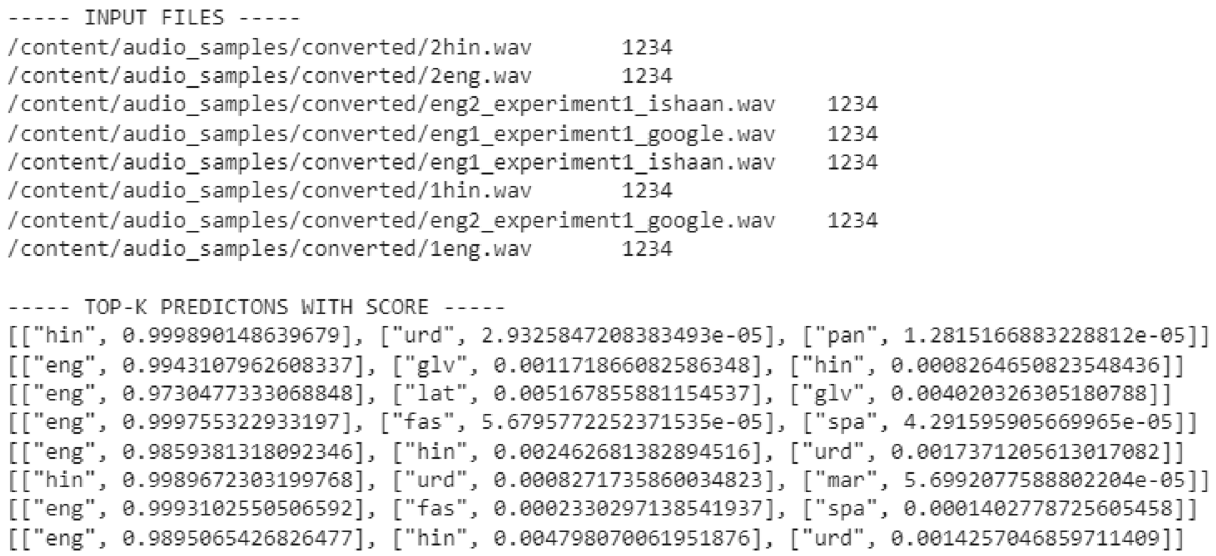
```
----- INPUT FILES -----
/content/audio_samples/converted/1hin_cropped.wav       1234
/content/audio_samples/converted/1eng_cropped.wav       1234
/content/audio_samples/converted/2hin_cropped.wav       1234
/content/audio_samples/converted/2eng_cropped.wav       1234

----- TOP-K PREDICTONS WITH SCORE -----
[["hin", 0.9932485818862915], ["urd", 0.005435482133179903], ["mar", 0.000582613458391279]]
[["eng", 0.870598316192627], ["kan", 0.039448246359825134], ["hin", 0.02283794991672039]]
[["hin", 0.9997424483299255], ["san", 6.999270408414304e-05], ["mar", 5.366690311348066e-05]]
[["eng", 0.9650427103042603], ["hin", 0.01581401377916336], ["urd", 0.006869655102491379]]
```

Recording	Length	Confidence score
1hin.wav	0:19	0.9989
2hin.wav	0:19	0.9998
1eng.wav	0:16	0.9895
2eng.wav	0:14	0.9943

Recording	Length	Confidence score
1hin_cropped.wav	0:06	0.9989
2hin_cropped.wav	0:06	0.9998
1eng_cropped.wav	0:06	0.9895
2eng_cropped.wav	0:06	0.9943

Experiment 2: Variation of scores with accent

To verify the hypothesis that the lower confidence scores for english recordings were due to my local accent, I recorded the same sentences being spoken in a native english accent and compared confidence scores.



```
----- INPUT FILES -----
/content/audio_samples/converted/2hin.wav       1234
/content/audio_samples/converted/2eng.wav       1234
/content/audio_samples/converted/eng2_experiment1_ishaan.wav    1234
/content/audio_samples/converted/eng1_experiment1_google.wav    1234
/content/audio_samples/converted/eng1_experiment1_ishaan.wav    1234
/content/audio_samples/converted/1hin.wav       1234
/content/audio_samples/converted/eng2_experiment1_google.wav    1234
/content/audio_samples/converted/1eng.wav       1234

----- TOP-K PREDICTONS WITH SCORE -----
[["hin", 0.999890148639679], ["urd", 2.9325847208383493e-05], ["pan", 1.2815166883228812e-05]]
[["eng", 0.9943107962608337], ["glv", 0.001171866082586348], ["hin", 0.0008264650823548436]]
[["eng", 0.9730477333068848], ["lat", 0.005167855881154537], ["glv", 0.004020326305180788]]
[["eng", 0.999755322933197], ["fas", 5.6795772252371535e-05], ["spa", 4.291595905669965e-05]]
[["eng", 0.9859381318092346], ["hin", 0.002462681382894516], ["urd", 0.0017371205613017082]]
[["hin", 0.9989672303199768], ["urd", 0.0008271735860034823], ["mar", 5.6992077588802204e-05]]
[["eng", 0.9993102550506592], ["fas", 0.0002330297138541937], ["spa", 0.0001402778725605458]]
[["eng", 0.9895065426826477], ["hin", 0.004798070061951876], ["urd", 0.001425704685971140 9]]
```

Task 2: MMS-TTS models - speech generation [English, Native]

For this task I use the `mms-tts-eng` and `mms-tts-hin` models from 🤗 HuggingFace Transformers. I run the TTS to generate sample audio from all four recordings:

- 1eng_tts.wav
- 2eng_tts.wav
- 1hin_tts.wav
- 2hin_tts.wav

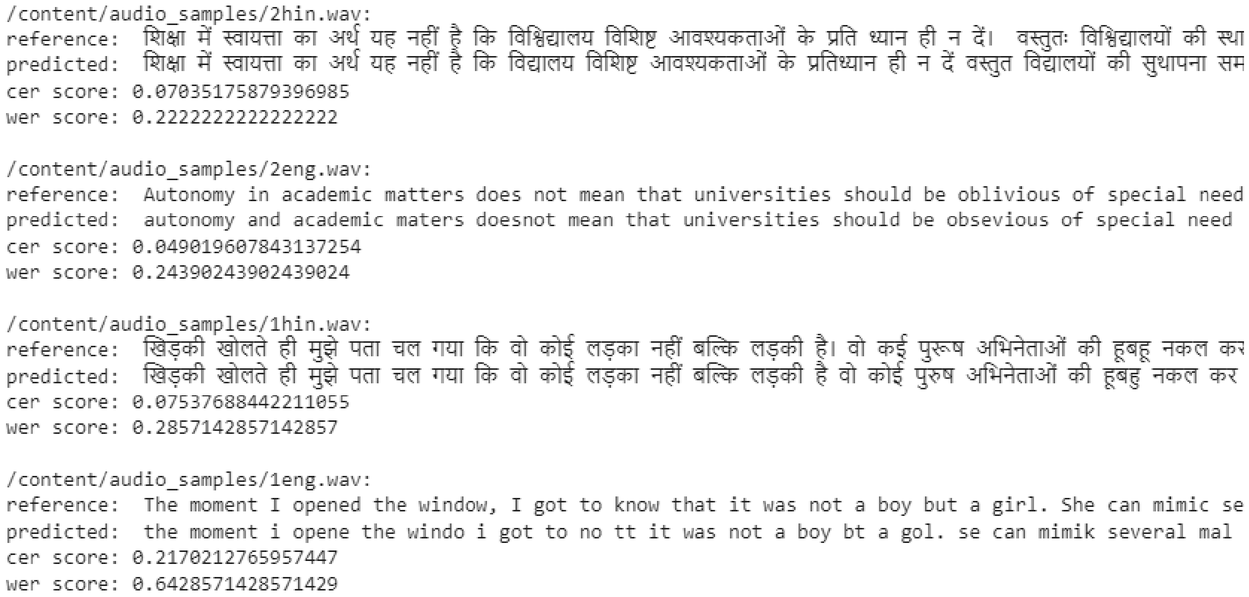
These recordings were generated on the corresponding paragraphs earlier mentioned in the report. You can find them in the zip file submitted alongside the report.

Task 3: MMS-ASR models - task1 transcription performance (CER, WER for recording vs generated audio transcriptions, comparison between English and Native)

I have used the `mms-1b-fl102` model which can run ASR on 102 language. The output language can be controlled by simply changing the language adapter on the model rather than changing all of the weights which is what makes this model convenient to use and efficient for applications.

I have generated transcriptions on the following 4 recordings:

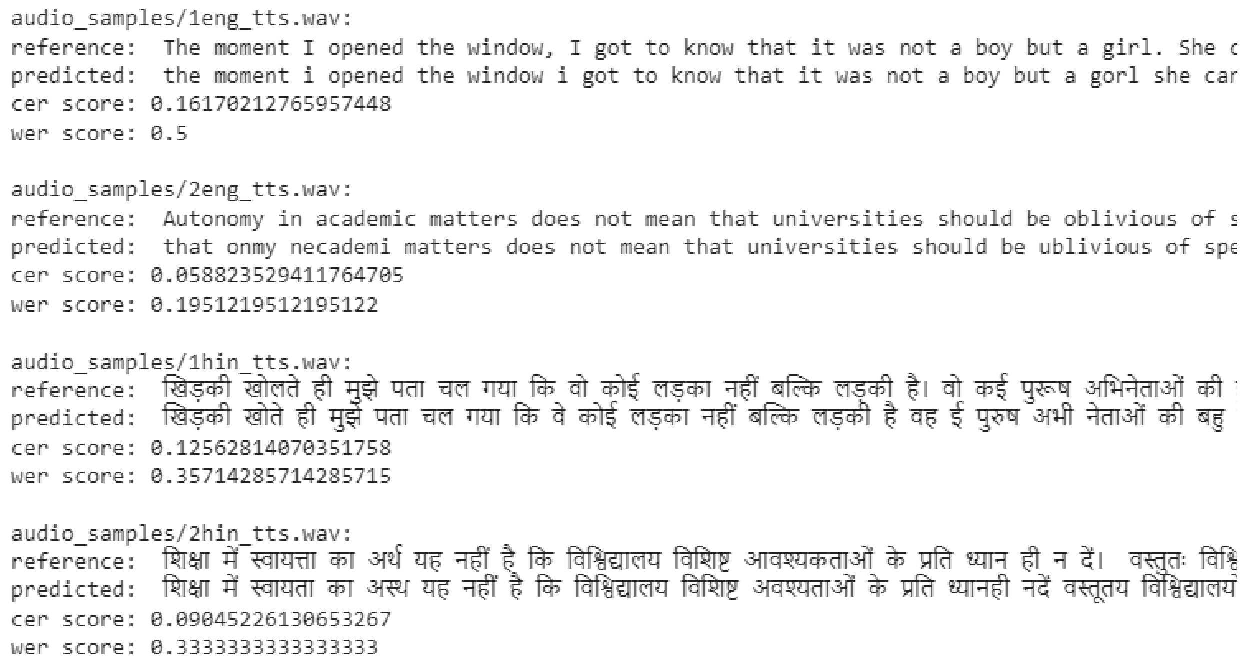
- 1eng.wav
- 2eng.wav
- 1hin.wav
- 2hin.wav



```
/content/audio_samples/2hin.wav:
reference:  शिक्षा में स्वायत्ता का अर्थ यह नहीं है कि विश्विद्यालय विशिष्ट आवश्यकताओं के प्रति ध्यान ही न दें।  वस्तुतः विश्विद्यालयों की स्था
predicted:  शिक्षा में स्वायत्ता का अर्थ यह नहीं है कि विद्यालय विशिष्ट आवश्यकताओं के प्रतिध्यान ही न दें वस्तुत विद्यालयों की सुथापना सम
cer score: 0.07035175879396985
wer score: 0.2222222222222222

/content/audio_samples/2eng.wav:
reference:  Autonomy in academic matters does not mean that universities should be oblivious of special need
predicted:  autonomy and academic maters doesnot mean that universities should be obsevious of special need
cer score: 0.049019607843137254
wer score: 0.24390243902439024

/content/audio_samples/1hin.wav:
reference:  खिड़की खोलते ही मुझे पता चल गया कि वो कोई लड़का नहीं बल्कि लड़की है। वो कई पुरूष अभिनेताओं की हूबहू नकल कर
predicted:  खिड़की खोलते ही मुझे पता चल गया कि वो कोई लड़का नहीं बल्कि लड़की है वो कोई पुरुष अभिनेताओं की हूबहु नकल कर
cer score: 0.07537688442211055
wer score: 0.2857142857142857

/content/audio_samples/1eng.wav:
reference:  The moment I opened the window, I got to know that it was not a boy but a girl. She can mimic se
predicted:  the moment i opene the windo i got to no tt it was not a boy bt a gol. se can mimik several mal
cer score: 0.2170212765957447
wer score: 0.6428571428571429
```

I have also generated transcriptions from the TTS versions of these recordings generated from their ground-truth sentences using earlier methods:

- 1eng_tts.wav
- 2eng_tts.wav
- 1hin_tts.wav
- 2hin_tts.wav



```
audio_samples/1eng_tts.wav:
reference:  The moment I opened the window, I got to know that it was not a boy but a girl. She c
predicted:  the moment i opened the window i got to know that it was not a boy but a gorl she car
cer score: 0.16170212765957448
wer score: 0.5

audio_samples/2eng_tts.wav:
reference:  Autonomy in academic matters does not mean that universities should be oblivious of s
predicted:  that onmy necademi matters does not mean that universities should be ublivious of spe
cer score: 0.058823529411764705
wer score: 0.1951219512195122

audio_samples/1hin_tts.wav:
reference:  खिड़की खोलते ही मुझे पता चल गया कि वो कोई लड़का नहीं बल्कि लड़की है। वो कई पुरूष अभिनेताओं की
predicted:  खिड़की खोते ही मुझे पता चल गया कि वे कोई लड़का नहीं बल्कि लड़की है वह ई पुरुष अभी नेताओं की बहु
cer score: 0.12562814070351758
wer score: 0.35714285714285715

audio_samples/2hin_tts.wav:
reference:  शिक्षा में स्वायत्ता का अर्थ यह नहीं है कि विश्विद्यालय विशिष्ट आवश्यकताओं के प्रति ध्यान ही न दें।  वस्तुतः विश्वि
predicted:  शिक्षा में स्वायता का अस्थ यह नहीं है कि विश्विद्यालय विशिष्ट अवश्यताओं के प्रति ध्यानही नदें वस्तूतय विश्विद्यालय
cer score: 0.09045226130653267
wer score: 0.3333333333333333
```

Arranging the scores into a table for comparison and analysis:

Recording	**Language**	**CER score**	**WER score**
1hin.wav	Hindi	**0.0753**	**0.2857**
2hin.wav	Hindi	**0.0703**	**0.2222**
1eng.wav	English	0.2170	0.6428
2eng.wav	English	**0.0490**	0.2439

Recording	**Language**	**CER score**	**WER score**
1hin_tts.wav	Hindi	0.1256	0.3571
2hin_tts.wav	Hindi	0.0904	0.3333
1eng_tts.wav	English	**0.1617**	**0.5000**
2eng_tts.wav	English	0.0588	**0.1951**

The better recording across the two of TTS and Normal have been highlighted in bold.

It is evident that on the hindi recordings, the TTS transcriptions were worse than the spoken hindi recordings. This can be seen in the higher CER and WER scores across those two rows in each table. This could possibly be due to the native hindi accent I have being better than the TTS hindi in terms of closeness to training data distribution.

On the english recordings however, the TTS transcriptions had a slightly better performance than the english recordings. This can be seen in the slightly lower CER and WER scores across those two rows in each table. This could be because the TTS recordings had an accent that was closer to the training data distribution for that language, than my indian-english accent.

Comments

It would be interesting to run an experiment on an audio dataset where the performance of some of the above measures is clustered across demographics. This would be able to reveal model biases across notable societal divisions very effectively and would help in tackling this bias. However, the challenge is to find a dataset covering a large enough user base with a wide spread across its own demographic groups as well as to maintain the privacy of the individuals involved.

References:

https://console.cloud.google.com/vertex-ai/generative/speech/text-to-speech?project=psyched-silicon-394610

https://huggingface.co/docs/transformers/model_doc/mms