

# Introduction to Audio Applications

## Assessment A1 Speech Processing

**Goal: The aim is to get familiarize with different audio-related Tasks.**

Note: All the artifacts of the assignments can be found [here](#), please use the institute id to access them.

### Task 1: Language Identification (LID)

#### Code summary:

- Clones fairseq-py repo and installs the latest version
- Downloads pre-trained MMS-LID model (l126) which can recognize 126 languages  
[List of 126 languages with code](#)
- Creates a folder for audio samples and prepares manifest files pointing to audio paths and durations  
Note:- audio files must be in .wav format with a sample rate of 16k.
- Set up environment variables like PYTHONPATH, Hydra configs, etc.
- Runs inference using the pre-trained MMS-LID model on the audio samples
- Inputs are the audio files
- Outputs top-k predictions for each audio with scores in the predictions.txt file

#### Results:

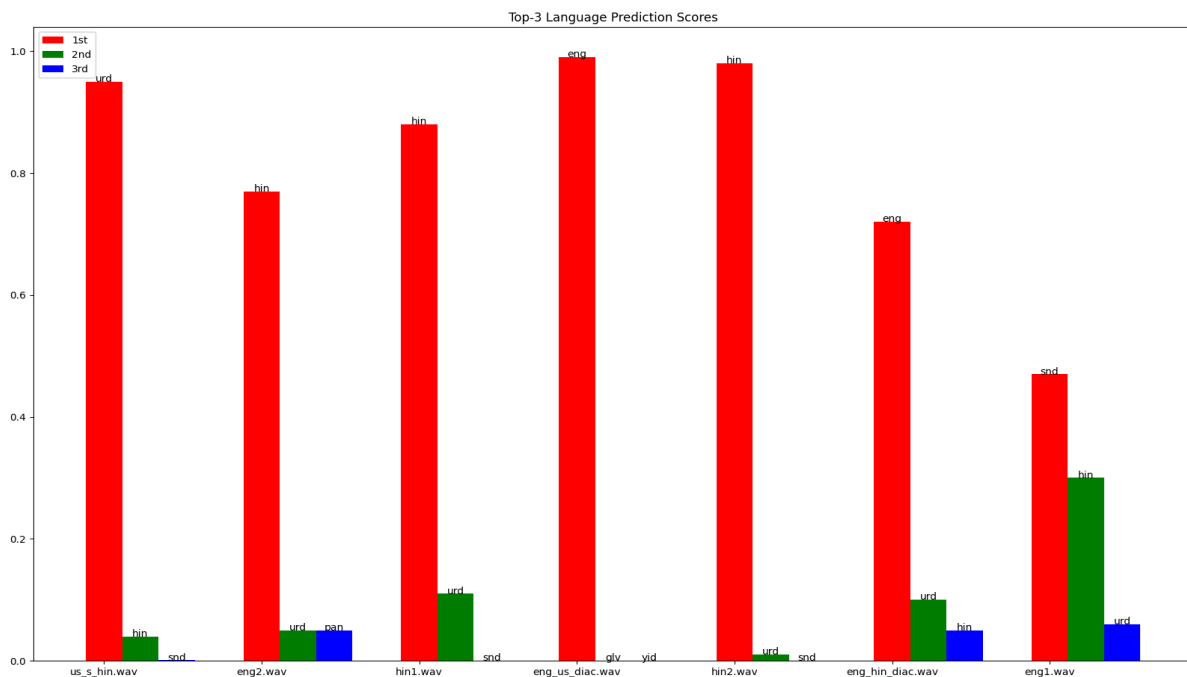


Fig 1: Plots of Language Identification Predictions

Where:

File Name	Description	Predictions top-3
eng1.wav	Recorded 1st sample in English	["snd",0.47874] ["hin",0.30364] ["urd",0.06962]
eng2.wav	Recorded 2nd sample in English	["hin", 0.77373] ["urd",0.05905] ["pan",0.05135]
hin1.wav	Recorded 1st sample in Hindi	["hin", 0.88729] ["urd",0.11074] ["snd",0.00071]
hin2.wav	Recorded 2nd sample in Hindi	["hin", 0.98696] ["urd", 0.01202] ["snd",0.00021]
eng_us_diac.wav	An American Person Speaking English	["eng", 0.99888] ["glv", 0.00044] ["yid", 0.00019]
eng_hin_diac.wav	An Indian Person Speaking English	["eng",0.72345], ["urd",0.10945] ["hin", 0.05807]
us_s_hin.wav	An American Person Speaking Hindi	["urd", 0.95206] ["hin",0.04718] ["snd", 9.16949e-05]]

Table 1: Table explaining the results with top-k predictions where k is 3

#### Analysis:

- The model is unable to correctly identify recorded voices.
- If the speaker is speaking at a slow pace i.e. taking more time to speak the next word, the model is getting confused.
- When an American person speaks English the model is identifying correctly with very high confidence
- When an Indian person speaks English the model confidence dops up to 0.72.
- Signifying the difference between dialects does affect the models' performance.
- Also when a non-Hindi speaker speaks Hindi, the model is identified as Urdu.

## Task 2: Text-to-Speech (TTS)

### Code summary:

- Clones VITS TTS repo and installs dependencies like PyTorch, Numpy, etc.
- Downloads pre-trained MMS TTS model checkpoint for a given language (English & Hindi).
- Loads model configuration, vocabulary, instantiate TTS model, load trained weights.
- Defines text processing - clean text, filter OOV words, convert to sequence.
- Preprocesses input text - lowercasing, cleaning, etc then convert to ID sequence.
- Feeds input text sequence to TTS model to generate mel spectrogram.
- Converts mel spectrogram to waveform reconstruction to output.

### Results:





File Name	Text	TTS
eng1	The sun shines brightly in the clear azure sky, casting its golden rays upon the earth, illuminating the world with its radiant warmth and splendor.	 gen_eng1.wav
eng2	The old oak tree stood majestically in the heart of the forest, its gnarled branches reaching towards the sky, providing shelter to a myriad of creatures that called the woods their home.	 gen_eng2.wav
hin1	साफ नीले आसमान में सूरज चमकता है, अपनी सोने जैसी किरणों को धरती पर गिराते हुए, अपनी तेज गर्मी और शान से दुनिया को रोशनी में ले जाता है।	 gen_hin1.wav
hin2	पुराना बलूत का पेड़ जंगल के दिल में शानदार रूप में खड़ा था, उसके उलझे हुए डालों ने आकाश की ओर बढ़ते हुए, जंगल के वो अनगिनत प्राणियों को आश्रय प्रदान किया जो जंगल को अपना घर कहते थे।	 gen_hin2.wav

Table 2: Results for TTS with audio links

**Analysis:**

- The model works with Multilingual inputs.
- Each Text to speech is given by a new voice actor, i.e. text is spoken by a different person even though the language is the same.
- The model suffers at “, its gnarled” in the eng2 text and can be heard in the output there can be several reasons for that one of them might be improper reconstruction from spectrogram to audio.
- Overall the speech generated by the model is very human-like and sounds natural.

**Task 3: Automatic Speech Recognition (ASR)**

**Code summary:**

- Clones fairseq repo, installs it. Downloads pre-trained MMS-ASR model checkpoint.
- Sets up environment variables and dependencies for inference.
- Prepares input audio files at 16KHz sample rate.
- Runs greedy decoding inference using the MMS model to transcribe audio to text.
- Calculate Word Error Rate and Character Error Rate metrics

**Result:**

Doc Index	Source	Prediction	Ground Truth
1	asr_eng 1.txt	te sun shines britly in de steer esu sky kasting ebes golden rave upon te ert illuminated te world with ex rediont vove and splonder	The sun shines brightly in the clear azure sky, casting its golden rays upon the earth, illuminating the world with its radiant warmth and splendor.
2	asr_eng 2.txt	ol ok tre stud majisticly in te hrt of de forest its gorned branches weching towards de sky providing shelder to mercles of crecrs dat coled te wod der hom	The old oak tree stood majestically in the heart of the forest, its gnarled branches reaching towards the sky, providing shelter to a myriad of creatures that called the woods their home.

3	asr_hin 1.txt	औफ़ नूले आसमान में सूरज चमकता है अपनी सोने जैसी किरणों को धरती पर गिराते हुए अपनी तेज़ गर्मी और शांत से दुनिया को रोशनी में ले जाता है	साफ नीले आसमान में सूरज चमकता है, अपनी सोने जैसी किरणों को धरती पर गिराते हुए, अपनी तेज गर्मी और शान से दुनिया को रोशनी में ले जाता है।
4	asr_hin 2.txt	र बलूत कापिल जंगल के दिल में शानदार रूपों में खाना था उसके ऊलझे हुए डालों में आकाश की ओर बढ़ते हुए जंगल के वह अनगिनत प्राणियों को आश्रय प्रदान किया जो जंगल को अपना घर कहते थे	पुराना बलूत का पेड़ जंगल के दिल में शानदार रूप में खड़ा था, उसके उलझे हुए डालों ने आकाश की ओर बढ़ते हुए, जंगल के वो अनगिनत प्राणियों को आश्रय प्रदान किया जो जंगल को अपना घर कहते थे।

Table 3: Model prediction for recoded voice

Doc Index	Source	Prediction	Ground Truth
1	asr_gen_en g1.txt	the sun shines brightly in the clear as your sky casting its golden rase upon the earth illuminating the world with its radiant warmth and splendor	The sun shines brightly in the clear azure sky, casting its golden rays upon the earth, illuminating the world with its radiant warmth and splendor.
2	asr_gen_en g2.txt	the oldupe tree stood magestically in the heart of the forest its maral branches reaching towards the sky providing shelter to a meriad of creatures that call the woods their home	The old oak tree stood majestically in the heart of the forest, its gnarled branches reaching towards the sky, providing shelter to a myriad of creatures that called the woods their home.
3	asr_gen_hin 1.txt	साफ़ नीय आसमान में सूरज चमकता है अपनी सोने जैसी किरणों को धरती पर गिराते हुए अपनी तेज गर्मी और शान से दुनिया को रोशनी में ले जाता है	साफ नीले आसमान में सूरज चमकता है, अपनी सोने जैसी किरणों को धरती पर गिराते हुए, अपनी तेज गर्मी और शान से दुनिया को रोशनी में ले जाता है।
4	asr_gen_hin 2.txt	पुराना बलूत का पेड़ जंगल के दिल में शानदार और रूप में खरा था उसके उलजे हुए डालों ने आकाश की ओर बढ़ते हुए जंगल के वो अनगिनत प्राणियों को आश्रय प्रदान किया जो जंकव को अपना घर कहते थे	पुराना बलूत का पेड़ जंगल के दिल में शानदार रूप में खड़ा था, उसके उलझे हुए डालों ने आकाश की ओर बढ़ते हुए, जंगल के वो अनगिनत प्राणियों को आश्रय प्रदान किया जो जंगल को अपना घर कहते थे।

Table 4: Model prediction for generated voice using MMS-TTS

Character Error Rate (CER) is a metric of the performance of an automatic speech recognition (ASR) system.

This value indicates the percentage of characters that were incorrectly predicted. The lower the value, the better the performance of the ASR system with a CharErrorRate of 0 being a perfect score. Character error rate can then be computed as:

$$CharErrorRate = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

where:

- $S$  is the number of substitutions,
- $D$  is the number of deletions,
- $I$  is the number of insertions,
- $C$  is the number of correct characters,
- $N$  is the number of characters in the reference ( $N=S+D+C$ ).

Fig 2: Formulation of Character Error Rate (CER)

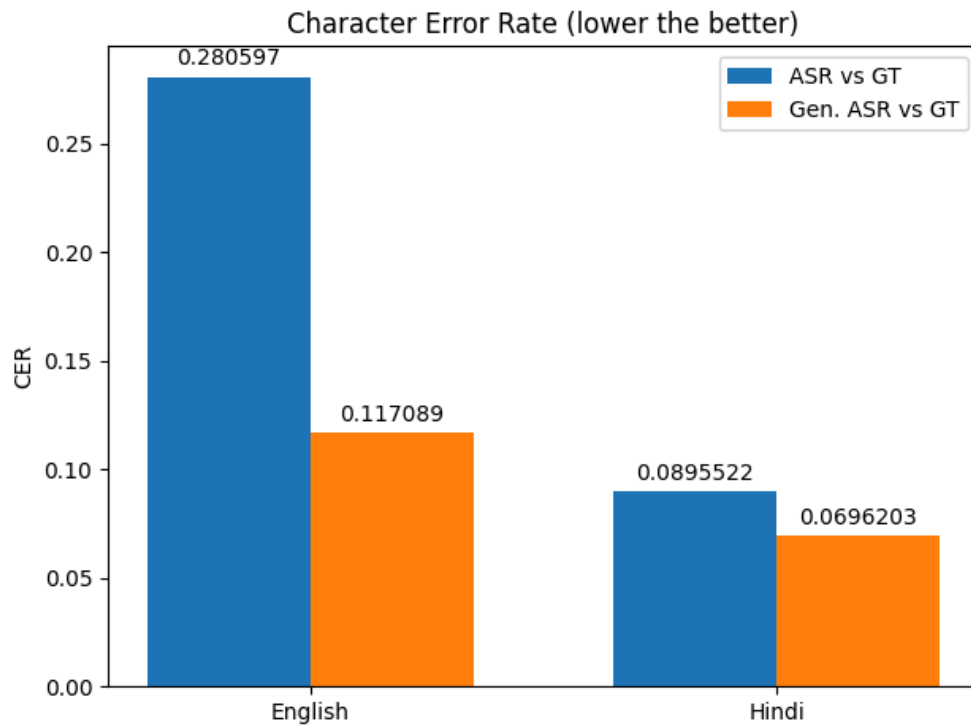


Fig 3: Character Error Rate between model prediction ASR from recorded audios and Ground Truth GT and Gen. ASR from the audios generated from the TTS task showing significant difference between CERs in English in comparison with Hindi.

Word error rate (**WordErrorRate**) is a common metric of the performance of an automatic speech recognition.

This value indicates the percentage of words that were incorrectly predicted. The lower the value, the better the performance of the ASR system with a WER of 0 being a perfect score. Word error rate can then be computed as:

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

where: -  $S$  is the number of substitutions, -  $D$  is the number of deletions, -  $I$  is the number of insertions, -  $C$  is the number of correct words, -  $N$  is the number of words in the reference ( $N = S + D + C$ ).

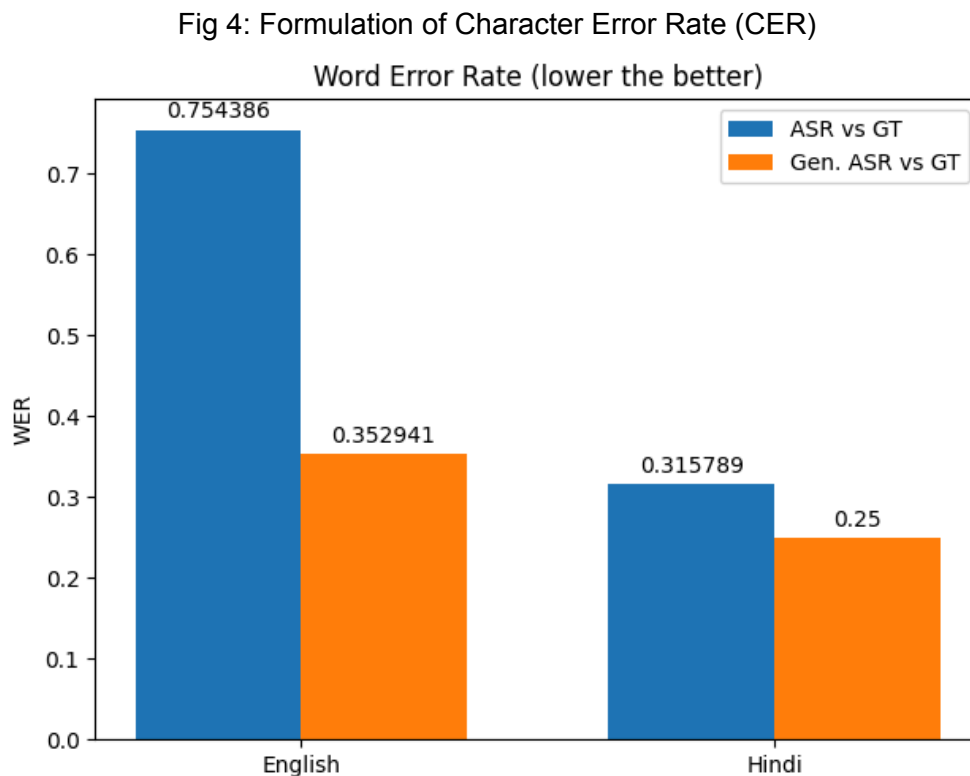


Fig 5: Word Error Rate between model prediction ASR from recorded audios and Ground Truth GT and Gen. ASR from the audios generated from the TTS task showing significant difference between CERs in English in comparison with Hindi.

Metric	Language	ASR vs GT	Gen. ASR vs GT
CER	English	0.2806	0.0896
CER	Hindi	0.1171	0.0696

WER	English	0.7544	0.3158
WER	Hindi	0.3529	0.2500

Table 5: Summarization of the results including CERs and WERs for the task of ASR.

#### Analysis:

- The model works with Multilingual inputs.
- ASR performance is on the shallow side in comparison to TTS performance due to the incorporation of factors like
  - Noise during recording
  - Speaking Pate / Speed of speech
  - Variation in Volume w.r.t to speaker's voice
  - Distance from Transducer (MIC)
  - Device Dependency
- The Results of ASR using TTS is very close to ground truth and can be seen in the metrics as well because of the eliminations of some of the above-mentioned factors that are not present in case of generations of audios and hence the audios are consistent.

#### References :

- [https://colab.research.google.com/github/facebookresearch/fairseq/blob/main/examples/mms/lid/tutorial/MMS\\_LID\\_Inference\\_Colab.ipynb](https://colab.research.google.com/github/facebookresearch/fairseq/blob/main/examples/mms/lid/tutorial/MMS_LID_Inference_Colab.ipynb)
- [https://colab.research.google.com/github/facebookresearch/fairseq/blob/main/examples/mms/tts/tutorial/MMS\\_TTS\\_Inference\\_Colab.ipynb](https://colab.research.google.com/github/facebookresearch/fairseq/blob/main/examples/mms/tts/tutorial/MMS_TTS_Inference_Colab.ipynb)
- [https://colab.research.google.com/github/facebookresearch/fairseq/blob/main/examples/mms/asr/tutorial/MMS\\_ASR\\_Inference\\_Colab.ipynb](https://colab.research.google.com/github/facebookresearch/fairseq/blob/main/examples/mms/asr/tutorial/MMS_ASR_Inference_Colab.ipynb)
- [https://lightning.ai/docs/torchmetrics/stable/text/char\\_error\\_rate.html](https://lightning.ai/docs/torchmetrics/stable/text/char_error_rate.html)
- [https://lightning.ai/docs/torchmetrics/stable/text/word\\_error\\_rate.html](https://lightning.ai/docs/torchmetrics/stable/text/word_error_rate.html)
- <https://github.com/pytorch/fairseq>