QUESTIONS

1. Read the Audio Files

a) What is the sampling rate of the provided audio files, and why is it important for speech recognition?

- All audio files in the dataset use a 16 kHz sampling rate.
- This is important because 16 kHz captures the full frequency range of human speech (0-8 kHz) while keeping audio size efficient. Most modern ASR models, including Whisper, are trained on and optimized for 16 kHz audio, so using this sampling rate improves recognition accuracy, clarity, and model compatibility.

b) What is the duration of each audio file?

Sr. No.	Utterance ID	Duration (sec)
1	121-127105-0000	9.875
2	121-127105-0001	5.025
3	121-127105-0002	7.495
4	121-127105-0003	7.725
5	121-127105-0004	2.110
6	121-127105-0005	5.820
7	121-127105-0006	4.725
8	121-127105-0007	5.790
9	121-127105-0008	2.760
10	121-127105-0009	2.290
11	121-127105-0010	2.850
12	121-127105-0011	5.780
13	121-127105-0012	4.830
14	121-127105-0013	5.895
15	121-127105-0014	2.255
16	121-127105-0015	2.960
17	121-127105-0016	2.030
18	121-127105-0017	2.695
19	121-127105-0018	2.770
20	121-127105-0019	3.525
21	121-127105-0020	14.355
22	121-127105-0021	2.000
23	121-127105-0022	5.075
24	121-127105-0023	10.910
25	121-127105-0024	14.450
26	121-127105-0025	16.065
27	121-127105-0026	7.530
28	121-127105-0027	13.870
29	121-127105-0028	6.750
30	121-127105-0029	7.310
31	121-127105-0030	2.175
32	121-127105-0031	10.765
33	121-127105-0032	3.170
34	121-127105-0033	2.355
35	121-127105-0034	7.410
36	121-127105-0035	14.150
37	121-127105-0036	4.150

c) What is the bit depth of the audio files, and how does it affect the quality of speech recognition?

Sr. No.	Utterance ID	Bit Depth
1	121-127105-0000	16
2	121-127105-0001	16
3	121-127105-0002	16
4	121-127105-0003	16
5	121-127105-0004	16
6	121-127105-0005	16
7	121-127105-0006	16
8	121-127105-0007	16
9	121-127105-0008	16
10	121-127105-0009	16
11	121-127105-0010	16
12	121-127105-0011	16
13	121-127105-0012	16
14	121-127105-0013	16
15	121-127105-0014	16
16	121-127105-0015	16
17	121-127105-0016	16
18	121-127105-0017	16
19	121-127105-0018	16
20	121-127105-0019	16
21	121-127105-0020	16
22	121-127105-0021	16
23	121-127105-0022	16
24	121-127105-0023	16
25	121-127105-0024	16
26	121-127105-0025	16
27	121-127105-0026	16
28	121-127105-0027	16
29	121-127105-0028	16
30	121-127105-0029	16
31	121-127105-0030	16
32	121-127105-0031	16
33	121-127105-0032	16
34	121-127105-0033	16
35	121-127105-0034	16
36	121-127105-0035	16
37	121-127105-0036	16

- Bit depth controls the dynamic range and precision of the audio signal.
- Higher bit depth means more accurate amplitude representation means clearer speech means fewer quantization errors.
- A 16-bit signal provides enough resolution for human speech and is the standard for ASR datasets**,** ensuring a good balance between quality and file size.

  If bit depth were too low (ex- 8-bit), the audio would have more noise and distortion, making ASR models like Whisper less accurate.

d) What is the file size of each audio file, and how might the size relate to audio quality or length?

Sr. No.	Utterance ID	File Size (KB)
1	121-127105-0000	177.60
2	121-127105-0001	102.02
3	121-127105-0002	142.21
4	121-127105-0003	137.91
5	121-127105-0004	34.21
6	121-127105-0005	109.36
7	121-127105-0006	86.06
8	121-127105-0007	108.22
9	121-127105-0008	52.83
10	121-127105-0009	52.35
11	121-127105-0010	59.72
12	121-127105-0011	114.76
13	121-127105-0012	96.71
14	121-127105-0013	111.93
15	121-127105-0014	40.64
16	121-127105-0015	58.57
17	121-127105-0016	45.20
18	121-127105-0017	58.75
19	121-127105-0018	58.31
20	121-127105-0019	68.95
21	121-127105-0020	272.83
22	121-127105-0021	45.56
23	121-127105-0022	88.84
24	121-127105-0023	209.78
25	121-127105-0024	268.38
26	121-127105-0025	301.52
27	121-127105-0026	143.75
28	121-127105-0027	257.05
29	121-127105-0028	124.62
30	121-127105-0029	133.57
31	121-127105-0030	42.41
32	121-127105-0031	173.29
33	121-127105-0032	62.91
34	121-127105-0033	43.00
35	121-127105-0034	119.96
36	121-127105-0035	205.99
37	121-127105-0036	70.33

- The file size of each audio file is mainly determined by:
- **Audio duration** - longer clips naturally produce larger files because more samples are stored.
- **Sampling rate (16 kHz)** - higher sampling rates generate more data per second.
- **Bit depth (16-bit)** - higher bit depth stores more detail per sample, increasing file size.
- Since all files share the same sampling rate and bit depth, differences in file size directly reflect differences in audio length, not quality.
- If sampling rate or bit depth were higher, files would be larger and potentially offer better clarity, which can improve ASR accuracy.

2. Text Preprocessing

a) If your transcript had misrecognized or misspelled words, how did you address that? Could spell-checking or correction be integrated into your pipeline?

- In my Whisper output, there were a few minor recognition errors (ex- substitutions or slightly altered word forms). I handled these through a text cleaning pipeline (lowercasing, punctuation normalization, filler removal, lemmatization), which already reduces noise and standardizes the output.
- If needed, spell checking can be added to the pipeline using tools like pyspellchecker or SymSpell. These can correct simple misspellings (ex- “governess - governance”), but I did not apply automated correction because:
  - Spell checkers may incorrectly “fix” valid words from this literary text.
  - Whisper errors are already minimal.
  - WER evaluation should not introduce extra corrections that alter model output.
- Spell correction could be integrated after tokenization, but for fair ASR evaluation, preprocessing should stay consistent and conservative.
- In my implementation I have added provisions for it too but have not used it.

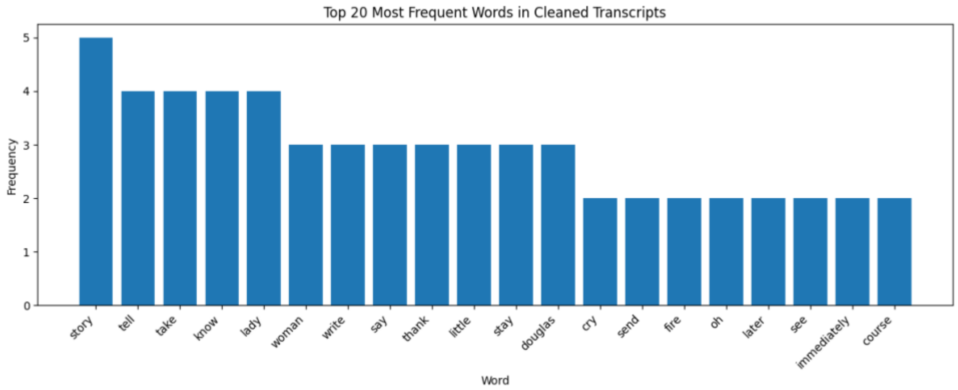
b) How would you modify your preprocessing pipeline if the transcript were multilingual or code-switched (i.e., contained multiple languages)?

- If the transcripts were multilingual or code switched, the preprocessing pipeline would need the following modifications:
  - Language detection step (ex- using langdetect or fastText) to identify which parts of the transcript belong to which language.
  - Switch to language appropriate tokenizers and models, such as spaCy’s multilingual model (xx_ent_wiki_sm), Stanza, or transformer based tokenizers, instead of using only the English en_core_web_sm.
  - Disable English only stopword removal, since removing English stopwords would incorrectly filter out meaningful words from other languages.
  - Avoid lemmatizing with a single language model and apply lemmatization separately for each detected language.
- These adjustments ensure that preprocessing remains accurate and fair when multiple languages appear in the same transcript.

3. EDA

○ Basic EDA

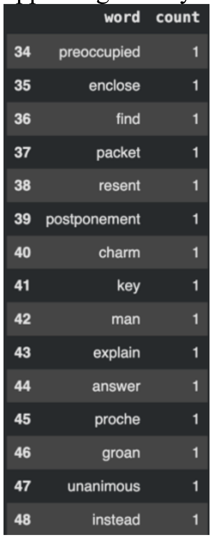
a) Visualize the top 20 most frequent words in the transcriptions. What do you observe?



- The most frequent words are story, tell, take, know, lady, woman, and other core content words.
- These words reflect the narrative and conversational nature of the dataset (a literary dialogue).
- Function words (the, and, I, etc.) are mostly absent because the cleaning pipeline removed stopwords and lemmatized remaining words.
- The distribution shows that the cleaned transcripts are content heavy, focusing on key nouns and verbs rather than filler or grammatical words.

b) Are there words that appear only once (hapax legomena)? What might they indicate?

- Number of hapax legomena (words appearing exactly once): 154

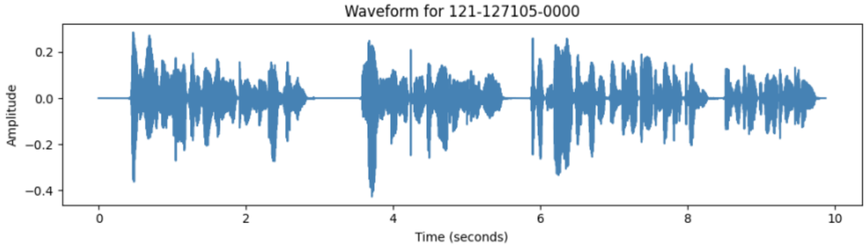


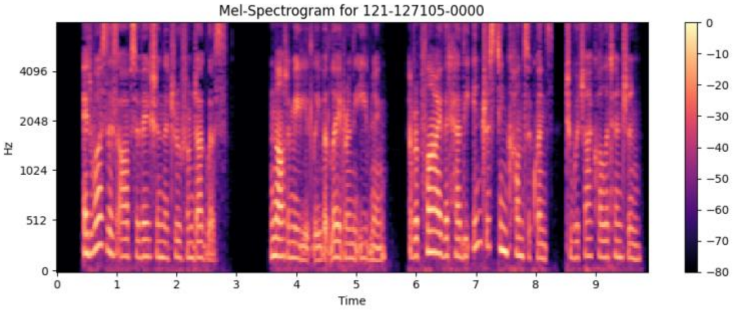
	word	count
34	preoccupied	1
35	enclose	1
36	find	1
37	packet	1
38	resent	1
39	postponement	1
40	charm	1
41	key	1
42	man	1
43	explain	1
44	answer	1
45	proche	1
46	groan	1
47	unanimous	1
48	instead	1

- These single occurrence words often indicate:
- Highly specific content words (ex- names, unique verbs, or uncommon adjectives).
- Narrative specific details that occur only in one sentence.
- Low frequency vocabulary that does not repeat across utterances.
- In NLP, a high number of hapax legomena typically reflects rich vocabulary variety, but it also makes tasks like language modeling and feature extraction sparser.

○ Audio-Specific EDA

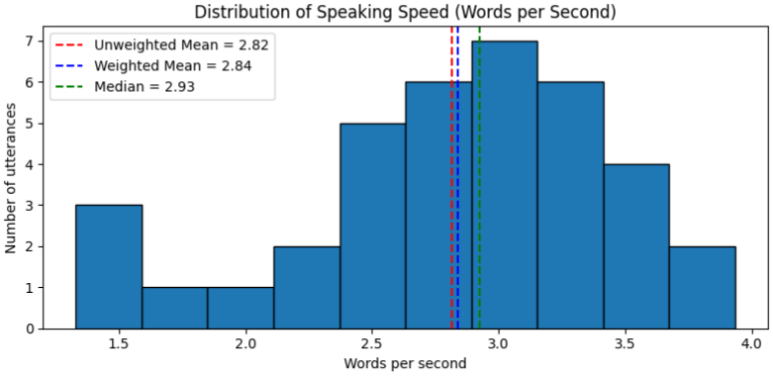
a) Plot a waveform or spectrogram of one audio file. What do you observe in terms of intensity or frequency distribution?

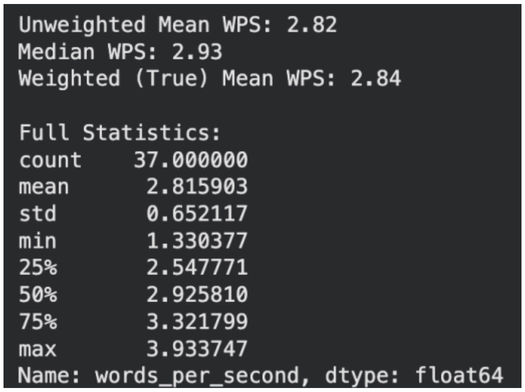




- The waveform plots show typical speech patterns, with noticeable bursts of energy during spoken segments and near silent pauses between phrases.
- The Mel-Spectrograms display clear formant structures and harmonic patterns characteristic of human speech.
- Dark vertical gaps represent natural pauses in the utterances.
- Brighter regions correspond to more intense sounds.
- The overall distribution confirms clean, high quality audio.

b) Is there a pattern in speaking speed (e.g., words per second)? Does this vary a lot across files?



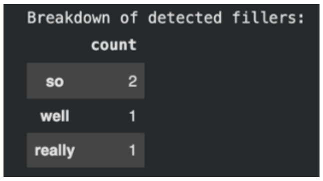


```
Unweighted Mean WPS: 2.82
Median WPS: 2.93
Weighted (True) Mean WPS: 2.84

Full Statistics:
count    37.000000
mean      2.815903
std       0.652117
min       1.330377
25%       2.547771
50%       2.925810
75%       3.321799
max       3.933747
Name: words_per_second, dtype: float64
```

- The speaking speed shows a moderate amount of variation across the 37 utterances. Most recordings fall in the range of 2.5-3.5 words per second, with a few slower lines (~1.3-2.0 WPS) and a few faster ones (~3.8-4.0 WPS).
- Overall, the pattern suggests:
  - Consistent mid range pace for most utterances.
  - Longer sentences tend to maintain a stable WPS, indicating the speaker reads at a controlled rate.
  - A few outliers appear slower or faster, likely due to sentence complexity or emphasis.
- In summary, speaking speed does vary, but not drastically, and the speaker remains mostly consistent across files.

c) Are there common filler words or disfluencies in the transcripts (e.g., "uh", "um", "you know")? Count and analyze.



Breakdown of detected fillers:

	count
so	2
well	1
really	1

- The dataset is clean, scripted speech, so fillers are naturally rare.
- The few fillers that appear are discourse connectors, not spontaneous hesitations (ex- “uh”, “um”).
- This suggests the speaker's delivery is fluent and structured, with minimal disfluency.
- For ASR evaluation, this means filler related recognition errors are minimal, keeping WER low.

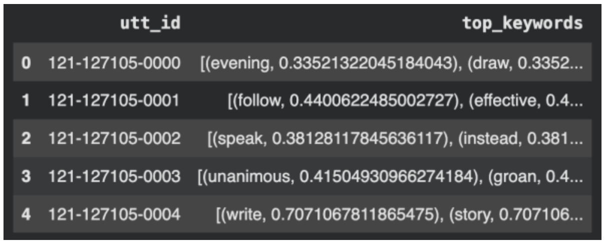
4. Feature Extraction

○ Text-Based Feature Extraction

a) What features can you extract from the text transcripts to represent them numerically (e.g., TF-IDF, bag-of-words, n-grams)? Use two techniques and compare your results.

- I represented the transcripts numerically using two techniques:
- Bag-of-Words (BoW): Converts text into raw word count vectors. It captures how often each word appears but ignores importance.
- TF-IDF: Weighs words by how unique or informative they are across transcripts. Highlights key terms and down weights common ones.
- Comparison:
  BoW gives simple frequency based features, while TF-IDF provides more meaningful, discriminative representations. Both matrices are sparse, but TF-IDF is more informative for analysis.

b) Can you identify keywords or phrases that are characteristic of certain speakers or topics in the transcripts?

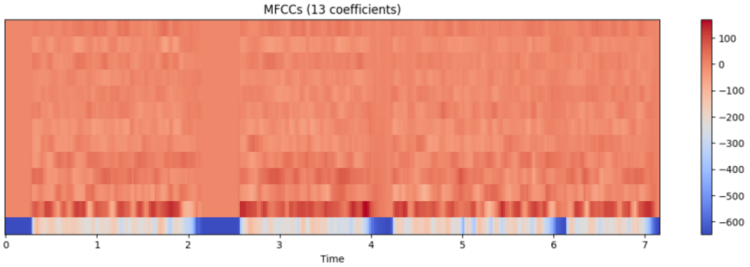


	utt_id	top_keywords
0	121-127105-0000	[(evening, 0.33521322045184043), (draw, 0.3352...
1	121-127105-0001	[(follow, 0.4400622485002727), (effective, 0.4...
2	121-127105-0002	[(speak, 0.38128117845636117), (instead, 0.381...
3	121-127105-0003	[(unanimous, 0.41504930966274184), (groan, 0.4...
4	121-127105-0004	[(write, 0.7071067811865475), (story, 0.707106...

- Using per utterance TF-IDF, each sample's top weighted terms were extracted as its most characteristic keywords.
- These keywords reflect the main idea or focus of that utterance.
- Since all utterances come from the same narrator/topic, the keywords mostly highlight local content differences rather than different speakers.
- Still, the per sample keywords clearly show distinctive phrases tied to specific narrative moments, even though the overall vocabulary overlaps heavily across the dataset.

Audio-Based Feature Extraction

a) What audio features could be extracted using MFCCs?



- From the MFCC analysis, the audio yields:
  - MFCCs (13 coefficients): Capture the spectral shape of speech and are widely used to represent vocal tract characteristics.
  - RMS Energy: Measures signal loudness over time, useful for detecting pauses, silence, and emphasis patterns.
  - Zero Crossing Rate (ZCR): Indicates how often the signal crosses zero, helping characterize voiced vs unvoiced regions and noise levels.
- These features together summarize the timbre, energy, and articulation patterns of the speaker, making them essential inputs for speech recognition and classification tasks.

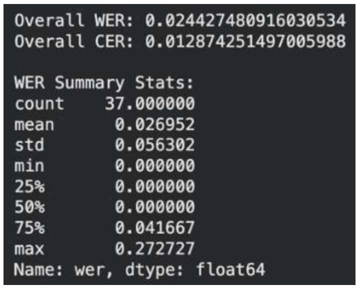
b) Would you use raw audio, features from ASR output, or both for downstream NLP tasks? Justify your choice.

- For most downstream NLP tasks, I would use ASR based textual features, because NLP models operate on text and benefit from clean, normalized transcripts (TF-IDF, n-grams, embeddings, etc.).
- However, for tasks that need emotion, speaker style, or acoustic cues, combining both is ideal. Raw audio features (MFCC, RMS, ZCR) add information that text alone cannot capture.

- Final Choice:
  - Text only for standard NLP tasks (classification, topic modeling, keyword extraction).
  - Audio + Text for richer tasks (emotion detection, speaker profiling, multimodal models).
- This gives the strongest overall performance depending on task requirements.

5. Evaluation

a) Word Error Rate (WER) is a standard metric in ASR that tells you how different your ASR output is from the reference transcript. Be sure to anlayze and discuss your response.



```
Overall WER: 0.024427480916030534
Overall CER: 0.012874251497005988

WER Summary Stats:
count    37.000000
mean      0.026952
std       0.056302
min       0.000000
25%       0.000000
50%       0.000000
75%       0.041667
max       0.272727
Name: wer, dtype: float64
```

- After applying a consistent normalization pipeline to both the Whisper predictions and the reference transcripts, I computed the overall transcription accuracy using Word Error Rate (WER) and Character Error Rate (CER). Whisper performed very well on this dataset, achieving an overall WER of 2.44% and CER of 1.29%, with half of all utterances reaching a perfect WER of 0. The error distribution was highly skewed toward zero, indicating that most transcriptions were accurate, with only a few utterances contributing small substitution or insertion errors. The highest WER (~27%) occurred in a single difficult utterance, but the majority remained below 4%. Overall, the analysis, implemented fully in my notebook assignment11_rajgorki.ipynb, shows that Whisper's mistakes were rare, isolated, and mostly limited to whole word substitutions rather than character level deviations.
- One limitation is that Whisper is trained on very large and diverse datasets, so this small and clean dataset may not fully reflect how the model performs on real world, noisier speech.