A comparative analysis of Automatic Speech Recognition Libraries

VASUDEVAN K

Abstract

This report compares the performance of three open-source speech recognition libraries: Vosk, Whisper, and wav2vec 2.0, in the context of commands for controlling an AFV. A single model from each library was evaluated on a dataset relevant to the application. This analysis focuses on the libraries' descriptions, RAM requirements, available models, supported languages, release dates, and transcription accuracy using the same dataset. The findings offer insights into the libraries' strengths and weaknesses, aiding readers in selecting a suitable option for their specific needs.

Introduction

Speech recognition technology plays a vital role in driving innovation and convenience across various applications. This report aims to compare and analyse three open-source speech recognition libraries: Vosk, Whisper, and wav2vec 2.0, to determine their suitability for this specific application.

While each library offers a diverse range of models, this evaluation focuses on a single, base model chosen from each library. This approach ensures a consistent and focused analysis allowing us to directly compare their performance on a dataset relevant to the project's application.

By evaluating these libraries based on their descriptions, resource requirements, available models, supported languages, and most importantly, transcription accuracy measured using the same project-specific dataset, this report will provide valuable insights into their strengths and weaknesses. Ultimately, this analysis aims to assist users in selecting the most effective speech recognition library for their specific needs, particularly those related to applications of this project.

Description of libraries

Vosk

Description

Vosk is an offline open-source speech recognition toolkit created by Alpha Cephei. Vosk models are small (50 Mb) but provide continuous large vocabulary transcription, zero-latency response with streaming API, reconfigurable vocabulary, and speaker identification. Vosk supplies speech recognition for chatbots, smart home appliances, virtual assistants. It can also create subtitles for movies, transcription for lectures and interviews. Vosk scales from small devices like Raspberry Pi or Android smartphone to big clusters. There are two types of models - big and small, small models are ideal for some limited tasks on mobile applications. They can run on smartphones, Raspberry Pi’s. They are also recommended for desktop applications. Big models are for the high-accuracy transcription on the server.

Methodology

Vosk relies on Hidden Markov Models, which are probabilistic models that represent sequences of states. Each state corresponds to a specific sound unit in the speech, and the transitions between states capture the probability of one sound following another.

Programming Languages supported - Python, Java, Node.JS, C#, C++, Rust, Go and others.

Languages Supported

English, Indian English, German, French, Spanish, Portuguese, Chinese, Russian, Turkish, Vietnamese, Italian, Dutch, Catalan, Arabic, Greek, Farsi, Filipino, Ukrainian, Kazakh, Swedish, Japanese, Esperanto, Hindi, Czech, Polish.

RAM requirements

Small model typically is around 50Mb in size and requires about **300Mb** of memory in runtime. Big models require up to **16Gb** in memory since they apply advanced AI algorithms.

Description of Models

Some of English Models are listed below (model used for evaluation is highlighted),

Model	**Size**	**Description**
vosk-model-small-en-us-0.15	40 Mb	Lightweight wideband model for Android and RPi
vosk-model-en-us-0.22-lgraph	128 Mb	Big US English model with dynamic graph
vosk-model-en-us-0.22	1.8 Gb	Accurate generic US English model
vosk-model-small-en-in-0.4	**36 Mb**	**Lightweight Indian English model for mobile applications**
vosk-model-en-in-0.5	1 Gb	Generic Indian English model for telecom and broadcast

Whisper

Description

OpenAI's Whisper is an automatic speech recognition (ASR) system trained on 680,000 hours of multilingual and multitask supervised data collected from the web. It shows that the use of such a large and diverse dataset leads to improved robustness to accents, background noise and technical language. Moreover, it enables transcription in multiple languages, as well as translation from those languages into English. We are open-sourcing models and inference code to serve as a foundation for building useful applications and for further research on robust speech processing. Whisper is robust to background noise and performs best at SNR ≤ 5dB.

Methodology

The Whisper architecture is a simple end-to-end approach, implemented as an encoder-decoder Transformer. Input audio is split into 30-second chunks, converted into a log-Mel spectrogram, and then passed into an encoder. A decoder is trained to predict the corresponding text caption, intermixed with special tokens that direct the single model to perform tasks such as language identification, phrase-level timestamps, multilingual speech transcription, and to-English speech translation.

Languages supported

Whisper does support transcription and translation across 98 languages and it performs best when sticking with English. The large model has a word error rate (WER) of 0.12 for English, 0.18 for Spanish,

0.23 for French, 0.25 for German and 0.28 for Mandarin2. However, some lower covered languages have much higher WER - e.g. Arabic (0.79), Hindi (0.86) and Swahili (1.00).

Description of Models (model used for evaluation is highlighted)

Size	Parameters	English-only model	Multilingual-model	VRAM needed	Speed (Relative)
Tiny	39 Mb	tiny.en	tiny	~ 1gb	32x
base	**74 Mb**	**base.en**	**base**	**~1gb**	**16x**
small	244 Mb	small.en	small	~ 2gb	6x
medium	769 Mb	medium.en	medium	~ 5gb	2x
large	1550 Mb	N/A	large	~ 10gb	1x

Wav2Vec 2.0

Description

wav2vec 2.0 is a complex and powerful speech recognition model developed by Facebook AI. It utilizes a self-supervised learning approach, meaning it learns from unlabelled audio data to capture essential speech features without requiring explicit transcriptions. This makes it potentially more efficient for training on large datasets and adapting to new domains with limited labelled data. Wav2vec trains models to learn the difference between original speech examples and modified versions, often repeating this task hundreds of times for each second of audio, and predicting the correct audio milliseconds into the future.

Methodology

Most current ASR models train on the log-Mel filter bank features of speech data, meaning audio that is been processed to make vocal features stand out. Wav2Vec turns raw speech examples into a representation — specifically, a code — that can be fed into an existing ASR system. The model then uses these representations to solve a self-supervised prediction task. Within each 10-second audio clip that the model is trained on, wav2vec generates several distractor examples, which swap out 10 ms of the original audio with sections from elsewhere in the clip. The model must then determine which version is correct. And this selection process is repeated multiple times for each 10-second training clip, essentially quizzing the model to discern accurate speech sounds from distractor samples hundreds of times per second.

Languages supported

It supports a total of 53 languages that includes Dutch, English, French, German, Italian, Polish, Portuguese, Spanish, Tamil, Hindi and so on.

RAM requirements

Meta has not explicitly mentioned the RAM usage of their models. Thus, with the help of psutil python package, an estimate of RAM usage of the model used for evaluation while transcribing a 15 second audio is obtained. It is approximated to be 500 – 600 Mb of RAM.

Description of Models (model used for evaluation is highlighted)

Model	Trained audio	Fine-tuned Audio	Language	Size
wav2vec2 - base - 960h	**960 hrs**	**960 hrs**	**English**	**378 Mb**
wav2vec2 - large - 960h	960 hrs	960 hrs	English	1.26 Gb

wav2vec2 - large - 960h - lv60 - self	53000 hrs	960 hrs + self-train	English	1.26 Gb
XLSR - 53 - Large	56000 hrs	N/A	Multilingual	3.5 Gb

EVALUATION OF MODELS

The models were tested with the list of commands that are related to the project's application. The following commands below were transcribed using the model and their performance is measured by calculating the Word Error Rate (WER) and Character Error Rate (CER). Making full use of **fastwer**, a python library, the WER and CER are easily calculated. (For accurate results the transcriptions are formatted to lower case, removed extra whitespaces, and removed special characters)

1. Turn left
2. Turn right
3. Turn right by [any angle]
4. Turn left by [any angle]
5. Brake
6. Turn on engine
7. Activate voice control
8. Deactivate voice control
9. Go forward
10. Go backward
11. Accelerate
12. Turn off engine
13. Turn on Engine
14. Adjust speed to [any number]
15. Help
16. Fire alert
17. Damage alert
18. Increase speed to [any number]
19. Decrease speed to [any number]
20. Lock hatch
21. Rotate turret by [any angle]

ANALYSIS 1

Using the above commands as inputs to model the following table is obtained. This table gives a view on individual performance of each model based on CER, WER and Execution time taken by model. The following obtained for inputs by single male voice in noiseless environment.

TABLE 1

Commands	Whisper			Vosk			Wav2Vec		
	WER	CER	Time(s)	WER	CER	Time(s)	WER	CER	Time(s)
1	0	0	1.27	0	0	0.36	100	67	0.3
2	0	0	0.42	0	0	0.35	50	50	0.25
3	0	0	0.46	60	67	0.65	100	75	0.37
4	0	0	0.43	60	55	0.6	60	64	0.28
5	100	40	0.39	100	40	0.37	100	20	0.21
6	0	0	0.42	33	50	0.3	33	7	0.24
7	0	0	0.39	0	0	0.37	67	23	0.31
8	0	0	0.43	67	13	0.52	100	25	0.28
9	0	0	0.38	0	0	0.38	100	50	0.23
10	0	0	0.4	0	0	0.37	0	0	0.26
11	0	0	0.39	0	0	0.38	0	0	0.22
12	0	0	0.42	67	80	0.36	67	20	0.24
13	33	7	0.41	33	7	0.32	67	29	0.23
14	43	38	0.47	86	59	0.63	71	35	0.33
15	0	0	0.37	0	0	0.28	0	0	0.2
16	0	0	0.37	0	0	0.37	50	20	0.22
17	100	33	0.4	50	50	0.56	100	25	0.21
18	43	38	0.47	43	18	0.6	71	28	0.28
19	57	44	0.48	43	23	0.69	100	59	0.33
20	0	0	1.43	50	60	0.24	0	0	0.23
21	0	0	0.48	80	56	0.56	140	93	0.38

The table 1 shows the results of evaluating three speech recognition models, Whisper, Vosk, and Wav2vec 2.0, on 21 commands. The evaluation metrics include Word Error Rate (WER), Character Error Rate (CER), and Execution Time.

Based on the table, Whisper achieves the best overall performance. It has the most instances (15 out of 21) where it perfectly transcribes the command (0% WER and CER). Vosk follows with 8 commands, and Wav2vec 2.0 lags behind with only 4 commands.

In terms of execution time, Whisper seems to be slower than both Vosk and Wav2vec 2.0 for most commands. However, it is important to consider the trade-off between accuracy and speed. Since Whisper achieves the highest accuracy, its slower execution time might be acceptable depending on the application.

Using the above table a graphical representation is created for comparing the WER, CER and Execution time of the three models considered for evaluation.

FIGURE 1.1

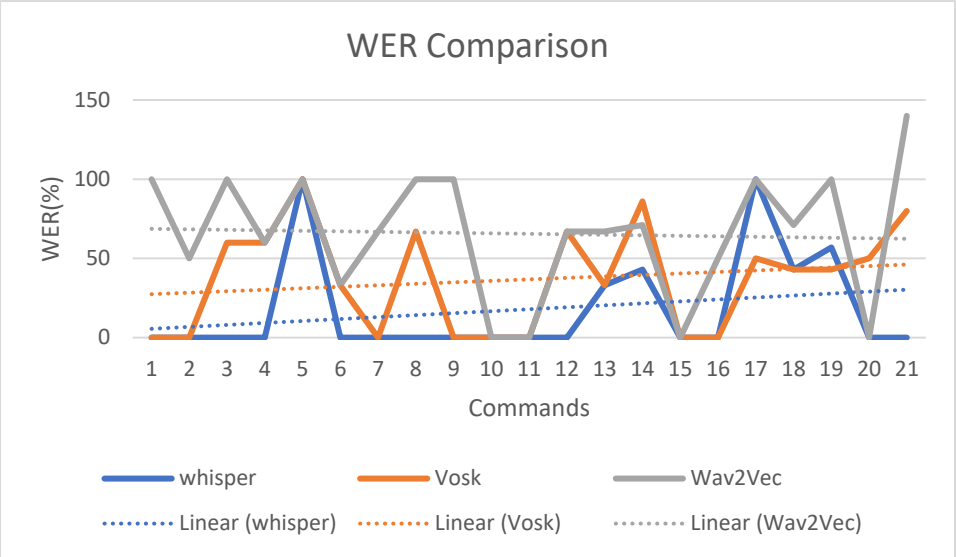


FIGURE 1.2

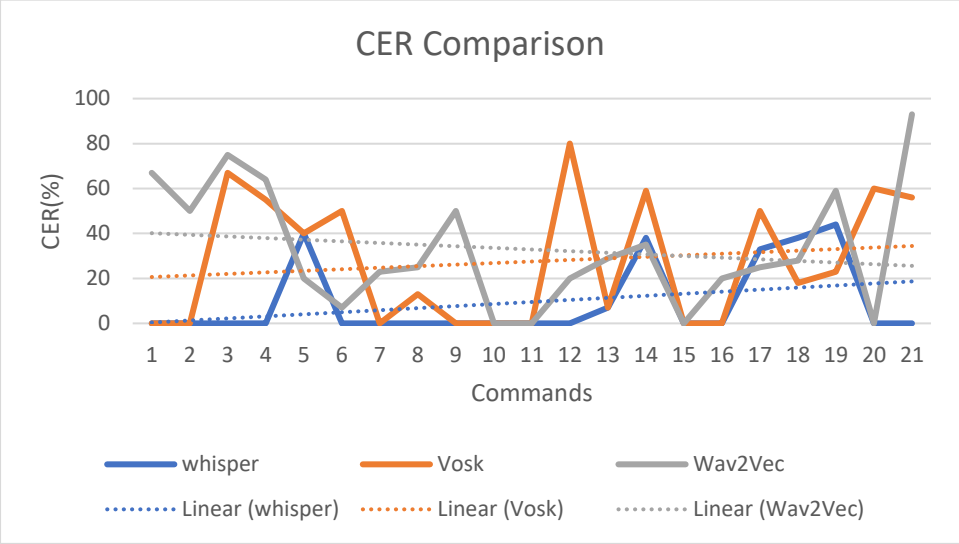
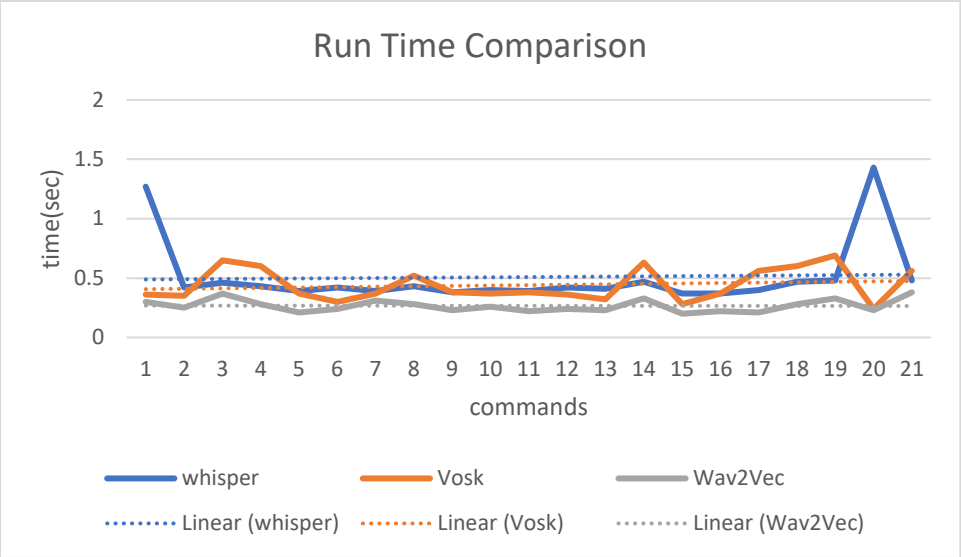


FIGURE 1.3



ANALYSIS 2.1:

COMPARISON OF MODELS OVER DIFFERENT VOICES USING CER

In this analysis two male voices (1 and 2) and a female voice (3) is used. Below table gives a view on performance of Whisper, Vosk and Wac2Vec models over three different voices compared with character error rate.

TABLE 2.1

commands	CER (%)								
	voice1			voice2			voice3		
	Whisper	Vosk	Wav2vec	Whisper	Vosk	Wav2vec	Whisper	Vosk	Wav2vec
1	0	0	66.67	0	100	33.33	0	66.67	66.67
2	0	0	50	0	50	70	0	90	40
3	0	66.67	75	0	100	94.12	0	43.48	100
4	0	54.55	63.64	0	93.75	75	0	69.57	104.35
5	40	40	20	60	100	40	40	20	40
6	0	50	7.14	57.14	35.71	71.43	0	14.29	64.29
7	0	0	22.73	36.36	63.64	72.73	0	13.64	59.09
8	0	12.5	25	4.17	33.33	50	0	12.5	54.17
9	0	0	50	30	100	30	0	0	70
10	0	0	0	27.27	27.27	36.36	0	9.09	27.27
11	0	0	0	0	0	40	140	80	90
12	0	80	20	0	46.67	66.67	26.67	0	60
13	7.14	7.14	28.57	7.14	35.71	35.71	7.14	92.86	50
14	37.84	59.46	35.14	0	72.22	72.22	27.27	63.64	90.91
15	0	0	0	0	50	0	75	100	75
16	0	0	20	0	80	30	0	90	70
17	33.33	50	25	50	75	50	83.33	33.33	50
18	38.46	17.95	28.21	0	85.71	90.48	0	58.33	62.5
19	43.59	23.08	58.97	10	55	75	0	83.33	75
20	0	60	0	20	10	0	80	40	80
21	0	55.56	92.59	20	80	65	0	42.31	69.23

Table 2.1. evaluates three speech recognition models (Whisper, Vosk, Wav2vec) on a set of commands spoken in three different voices (total of 63 commands when considering individual voices). The evaluation metric is character error rate (CER).

Whisper demonstrates superior performance. It achieves a 0% CER for an average of 60.3% of the commands across all voices. Additionally, 31.7% of commands have a CER below 50%. This stands in contrast to Vosk, which only achieves a 0% CER for an average of 17.5% of the commands, and Wav2vec, which has an even lower average of 9.5% commands with a 0% CER.

A graphical representation of comparison of model performance for different voices with CER.

FIGURE 2.1.1

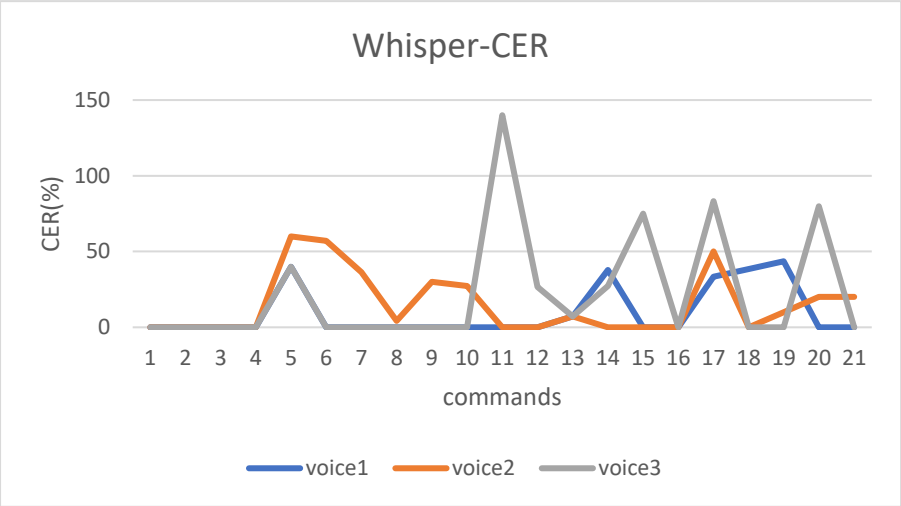


FIGURE 2.1.2

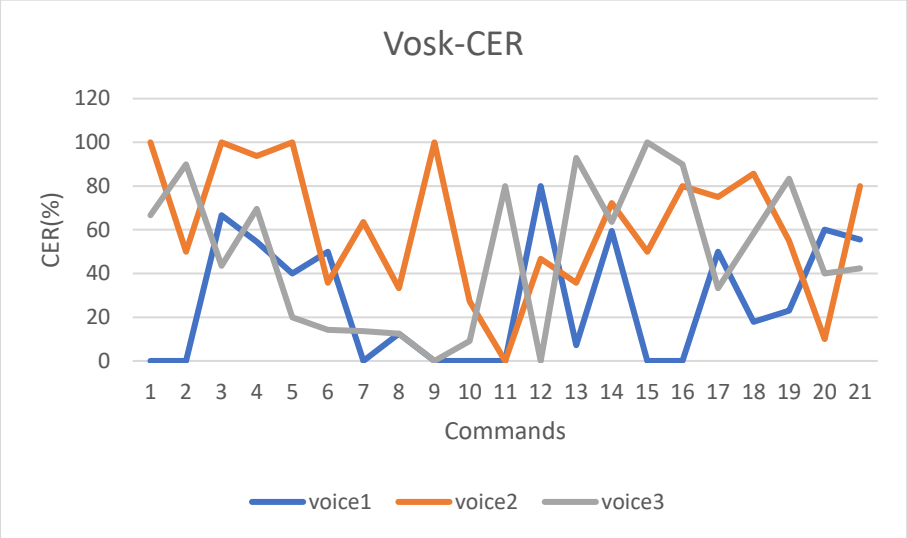
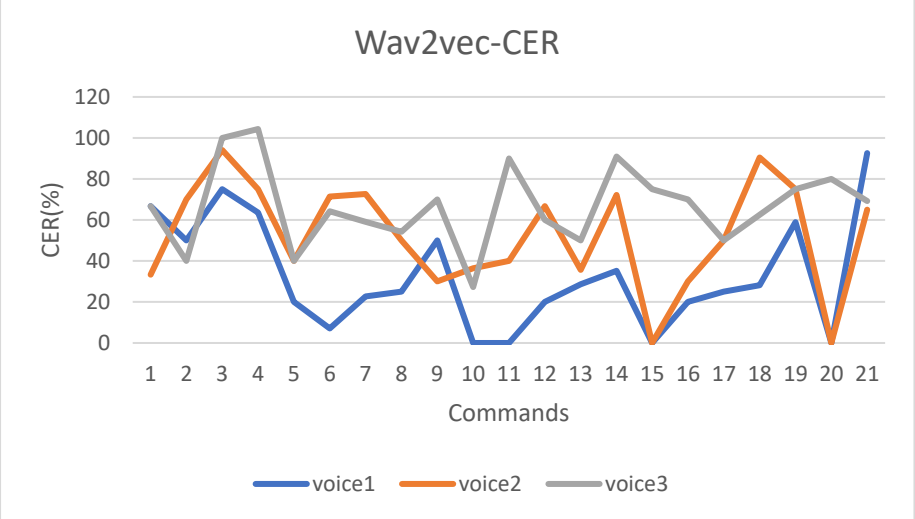


FIGURE 2.1.3



ANALYSIS 2.2: COMPARISON OF MODELS OVER DIFFERENT VOICES USING WER

In this analysis two male voices (1 and 2) and a female voice (3) is used. Below table gives a view on performance of Whisper, Vosk and Wac2Vec models over three different voices compared with word error rate.

TABLE 2.2

commands	WER (%)								
	voice1			voice2			voice3		
	Whisper	Vosk	Wav2vec	Whisper	Vosk	Wav2vec	Whisper	Vosk	Wav2vec
1	0	0	100	0	100	50	0	150	100
2	0	0	50	0	50	100	0	150	100
3	0	60	100	0	100	100	0	40	120
4	0	60	60	0	100	75	0	100	160
5	100	100	100	100	100	100	100	100	100
6	0	33.33	33.33	66.67	33.33	100	0	33.33	100
7	0	0	66.67	33.33	66.67	133.33	0	33.33	100
8	0	66.67	100	33.33	33.33	133.33	0	66.67	133.33
9	0	0	100	50	100	100	0	0	100
10	0	0	0	50	50	100	0	50	50
11	0	0	0	0	0	100	500	200	100
12	0	66.67	66.67	0	33.33	100	33.33	0	100
13	33.33	33.33	66.67	33.33	33.33	66.67	33.33	100	100
14	42.86	85.71	71.43	0	125	125	40	120	140
15	0	0	0	0	100	0	100	100	100
16	0	0	50	0	100	100	0	100	100
17	100	50	100	100	150	100	150	100	100
18	42.86	42.86	71.43	0	100	125	0	100	120
19	57.14	42.86	100	25	75	100	0	100	120
20	0	50	0	50	50	0	100	50	150
21	0	80	140	50	100	100	0	60	100

Table 2.2. evaluates three speech recognition models (Whisper, Vosk, Wav2vec) on a set of commands spoken in three different voices (total of 63 commands when considering individual voices). The evaluation metric is character error rate (WER).

Whisper demonstrates superior performance. It achieves a 0% WER for an average of 46% of the commands across all voices. Additionally, 14.3% of commands have a WER below 50%. This stands in contrast to Vosk, which only achieves a 0% WER for an average of 17.5% of the commands, and Wav2vec, which has an even lower average of 9.5% commands with a 0% WER.

A graphical representation of comparison of model performance for different voices with WER

FIGURE 2.2.1

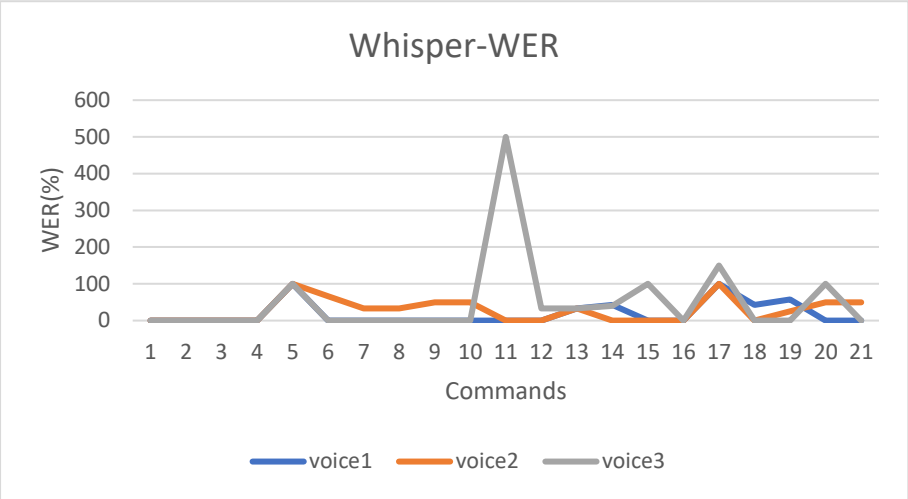


FIGURE 2.2.2

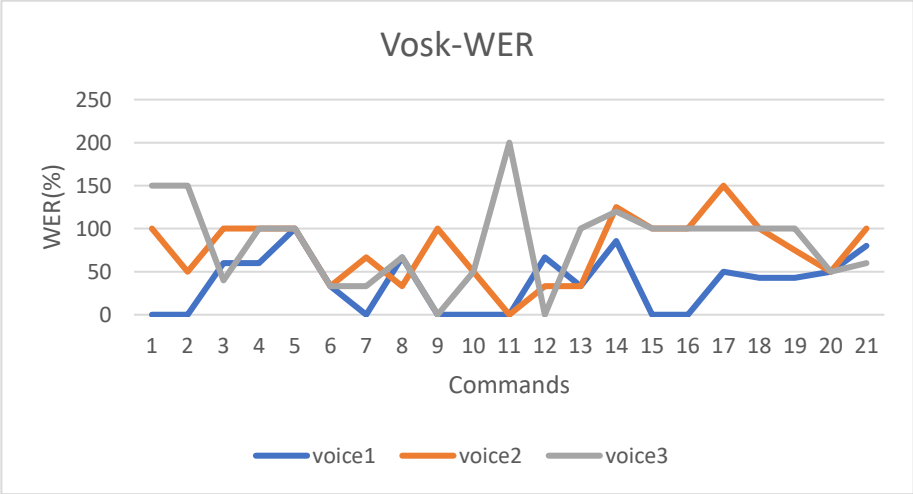
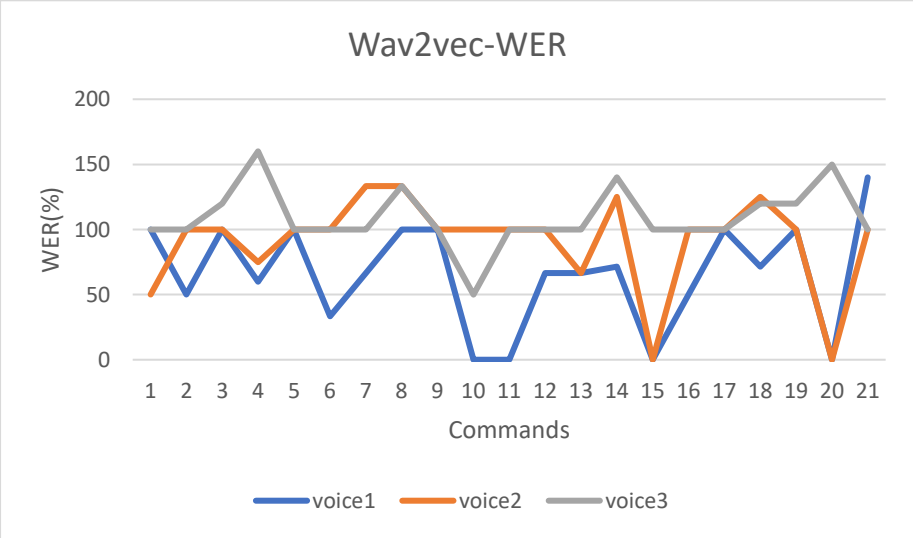


FIGURE 2.2.3



ANALYSIS 3: TESTING MODELS WITH A DATASET AND COMPARING THEIR PERFORMANCE WITH WER AND CER

Description of dataset – NPTEL2020 – Indian English speech dataset

GitHub Repo: https://github.com/AI4Bharat/NPTEL2020-Indian-English-Speech-Dataset

A Speech-to-Text dataset scraped from NPTEL for Indo-English accent, from Education Domain. The dataset is created from NPTEL lectures as source. It contains 15,700 hours of audio which are split into 6.2 million chunks of length 3 – 10 seconds.

The dataset is split as,

Split	**Number of Chunks**
Train set	5 M
Validation set	625 K
Test set	625 K
Sample set	1 K

Sample Set is a small subset manually annotated by AI4Bharat to compute the quality of data. It is also referred as Pure Set. This report has utilized the pure set for the evaluation of models. The models resulted in WER and CER for each chunk. They are finally averaged for 1k chunks and it is obtained in the following table.

Whisper		**Vosk**		**Wav2Vec**	
WER (%)	**CER (%)**	**WER (%)**	**CER (%)**	**WER (%)**	**CER (%)**
23.24	16.64	48.5	38.43	46.22	22.06

INFERENCE

Our evaluation across various metrics suggests Whisper as a strong contender for speech recognition tasks.

- **Accuracy:** As shown in Table 1 and the graphs (Analysis 1), the trendlines of Whisper consistently maintains a lower Word Error Rate (WER) and Character Error Rate (CER) compared to Vosk and Wav2Vec. This indicates Whisper's superior accuracy in recognizing spoken words.
- **Execution Time:** Notably, the execution time for each model across different commands exhibits negligible variation. This suggests similar processing speeds for all three models.
- **Voice Independence:** Interestingly, Figures 2.2.1 and 2.1.1 reveal that Whisper's performance remains relatively stable (lines tracking closely) for various voices. In contrast, Vosk and Wav2Vec show more variation in CER and WER across different speakers. This highlights Whisper's robustness to speaker variability.
- **Dataset:** the NPTEL dataset confirms Whisper's superiority in handling long sentences with minimal errors, making it the most suitable model for such tasks.
- **Error Rates:** As observed in all the tables, error rates greater than 100 denote that number of transcribed words are greater than the original text, which is insignificant.

Conclusion:

Taking all factors into account, particularly accuracy and voice independence, Whisper emerges as the most promising choice for general speech recognition tasks. Its competitive execution time further strengthens its position. Overall, Whisper's well-rounded performance across various metrics positions it as the recommended model for speech recognition applications.