Voice Data Collection and Pre-Processing Report

1. Objective

The objective of this activity was to collect authentic user voice queries related to the university admission system and prepare a clean, standardized audio dataset suitable for fine-tuning the Whisper Automatic Speech Recognition (ASR) model.

2. Data Collection Process

To gather real-world voice data, a comprehensive list of admission-related statements was prepared. These statements covered greetings, admission inquiries, programs offered, eligibility criteria, admission procedures, merit lists, fee structures, scholarships, academic schedules, facilities, and closing responses.

Participants were instructed to record the provided statements and optionally record any additional queries in English, Urdu, or mixed language, along with the written form of the query. The data collection session lasted approximately one hour. Despite initial difficulty convincing participants, a total of 1,376 voice recordings were successfully collected in .ogg format, commonly generated by mobile devices.

3. Challenges Faced

Limited interest from participants in recording voice samples, managing a large number of audio files, and variations in language, accent, and pronunciation were the main challenges. Despite these, the dataset was successfully gathered.

4. Audio Conversion and Renaming

To ensure compatibility with Whisper ASR, all .ogg files were converted to .wav format, standardized to mono channel, resampled to 16 kHz, and renamed using a consistent and anonymous naming scheme:

unk_mix_0001.wav, unk_mix_0002.wav, ...

Audio Conversion Command (PowerShell):

```
& "C:\Users\Home\Downloads\ffmpeg-2026-01-26-git-fe0813d6e2-full_build\ffmpeg-2026-01-26-git-fe0813d6e2-full_build\bin\ffmpeg.exe" `
```

```
-i *.ogg -ac 1 -ar 16000 wav\unk_mix_%04d.wav
```

Explanation: *.ogg → selects all .ogg files; -ac 1 → converts audio to mono; -ar 16000 → resamples to 16 kHz; wav\unk_mix_%04d.wav → stores files sequentially in wav folder.

5. Dataset Organization

The dataset was organized into Raw Dataset (.ogg recordings) and Processed Dataset (.wav recordings) to ensure reproducibility and preserve original data..

6. Transcription Process

After audio conversion and organization, the next step involved manually transcribing all collected voice recordings. This process ensured high-quality text corresponding to each audio sample, which is critical for supervised fine-tuning of the Whisper ASR model. The steps followed were:

6.1. Listening to Audio

- Each .wav file was played sequentially.
- Participants' mixed language (English, Urdu, or code-switched sentences) was carefully noted.

6.2. Manual Transcription

- The spoken content was typed verbatim in a plain text file (.txt).
- Special attention was given to punctuation, pauses, and proper nouns.
- Each transcription was saved using the same file naming convention as the audio (e.g., unk_mix_0001.txt).

6.3. Quality Check

- A second review of each transcription was conducted to correct misheard words, inconsistencies, or spelling errors.
- Any unclear or inaudible parts were marked with [inaudible] to maintain data integrity.

6.4. Structuring for CSV

- Once all transcriptions were completed, the audio-text pairs were compiled into a CSV file for Whisper fine-tuning.
- The CSV format was standardized as:

```
audio_filepath,text
wav/unk_mix_0001.wav,"Assalam o Alaikum, mujhe admission se related information chahiye"
wav/unk_mix_0002.wav,"Mujhe university admissions ke baare mein detail bata dein"
```

- This structured format ensures compatibility with Whisper’s training pipeline.

6.5. Language Labeling (Optional)

- Each entry was optionally labeled with the primary language (English, Urdu, Mix) to allow analysis of model performance on different language types.

7. Conclusion

This phase successfully produced a clean, structured, and Whisper-compatible voice dataset of 1,376 recordings, ready for transcription and fine-tuning.

8. Tools Used

FFmpeg (Windows build) was used for batch audio conversion, resampling, and channel standardization, essential for Whisper ASR fine-tuning.