MS(CS/SE) Research Thesis-I

Interim Report

Name	Manal Asghar
Student Id	FA24-MSDS-0004
Thesis Title	Data-Efficient Urdu Speech Corpus Construction
Supervisor Name	Dr. Tafseer Ahmed

List of Meetings

S. No.	Date	Topic
1.	10 Oct,2025	Discussion about the topic
2.	18 Oct,2025	Summary of research paper
3.	21 Oct,2025	Gap analysis and Novelty
4.	23 Oct,2025	Architecture
5.	1 Nov,2025	Interim report

List of Research Papers for Literature Review/ Gap Analysis (at least 5)

(Note that you must submit the papers along with this document)

S.No.	Paper Title	Journal/ Conference	Year
1.	A Survey on Data Selection for Efficient Speech Processing (2025)	Survey article (journal/conference)	**2025**
2.	Diversity-Based Core-Set Selection for Text-to-Speech with Linguistic and Acoustic Features (Seki et al.)	ICASSP 2024 (Seoul, Korea)	**2023 preprint 2024 (conference)**
3.	Enhancing Low-Resource ASR through Versatile TTS: Bridging the Data Gap (Yang et al.)	arXiv - submitted for conference (preprint), accepted for **(ICASSP)**	**2024**
4.	Speech Data Selection for Efficient ASR Fine-Tuning using Domain Classifier and Pseudo-Label Filtering (Rangappa et al.)	ICASSP	**2025**
5.	Unsupervised Active Learning: Optimizing Labeling Cost-Effectiveness for Automatic Speech Recognition (Zheng et al.)	Interspeech 2023 (Dublin, Ireland)	**2023**

Abstract

(750 words extended Abstract mainly focusing on research background, but also mentioning potential gap and solution(s).)

Automatic Speech Recognition and Text-to-Speech systems for low-resource languages like Urdu are also a significant challenge in the current research of speech technology. Whereas languages such as English also has massive and high-quality annotated speech databases Urdu is still under a critical resource constraint. Though much of the Urdu writing is found in the news, literature, and electronic media, a remarkable lack of parallel speech information is evident. This is also a major weakness of speech models and will not allow scalable improvements in the development of Urdu ASR and TTS. Annotation and collection of speech data are a time-consuming, expensive and manual task. Consequently, the majority of previous studies have also been based on small task-specific or simple data augmentation techniques instead of answering a critical question: how can we find and select the most valuable text to record speech in order to maximize the usefulness of the dataset and minimize the amount of annotation work?

The proposed study will fill that gap, by proposing a data and intelligence-saving framework of text selection to create a high-quality Urdu speech corpus using a limited number of resources. The offered approach points to the fact that all text samples do not affect the model learning to the same extent. The framework does not rely on random or heuristic sampling but on the selection of text to represent the linguistic diversity of Urdu in the best way possible. The process is inspired by the concepts of diversity-based and active learning and makes sure that the chosen sentences represent the rich linguistic diversity in terms of the number of styles, structures, and vocabulary patterns that are the basis of the natural use of the Urdu language. With the emphasis on representativeness, the framework ensure that the data obtained to aid in speech annotation is not only small but richly contented as well.

The Urdu text will be gathered through publicly available sources such as news and social media sites. These sources give natural variability in the use of language such as regional variations, stylistic variations, and varying degrees of formality. Once the data is collected, a selection algorithm will be used whereby each sentence is ranked according to diversity and linguistic richness.

After choosing the subset, native speakers of Urdu will record the text as a consistency measure and to make the text clear. Instead of creating a large corpus, this project proves that a small corpus with active records can perform equally to much bigger, randomly gathered corpus. The approach puts an emphasis on the data quality instead of data quantity.

The transcribed content will be evaluated with the help of modern multilingual transformer based ASR systems Whisper and XLS R, with the help of such established metrics as Word Error Rate (WER) and Character Error Rate (CER). This analysis will reveal whether a cautiously chosen and documented data can attain comparable or even better results at a reduced cost and effort. The trained vocabulary will also be useful in the development of TTS since the same curated set can be used as a single linguistic base of both recognition and synthesis.

This approach can be applied to other low resource South Asian languages other than Urdu like Sindhi and Balochi. It offers a feasible way of developing high quality speech resources without the need to invest heavily in financial and computational solutions. in general, such a study highlights the notion that with low resource speech technology, smart data is more important than not more data to spur real innovation in ASR and TTS.

Conclusively, the research paper propose a systematic and human-oriented method to constructing Urdu speech corpus by intelligent text selection and effective human recordings. This project is not only concerned with the

shortage of Urdu speech resources but also provides a basis to develop the methods of data development in the long run. The framework is practical and can be extended to other underrepresented languages to enable the betterment of a more inclusive future of speech technology; primarily it is dedicated to assisting low-resource language users in making AI speech tools more available, dependable, and culturally relevant.

Project Plan

(Planning and scheduling of your project in phases)

Phase	Date	Task
Research Phase	20-Nov	Literature Review
	1-Dec	Literature Review Writing
Planning Phase	15-Dec	Methodology
Data Preparation Phase	25-Dec	Text Corpus Collection
	15-Jan	Method 1: Text Subset Extraction
	15-Mar	Method 2: Text Subset Extraction
Modeling Phase	30-Apr	Speech Dataset & Model Training
Evaluation Phase	10-May	Performance Comparison
Finalization Phase	30-May	Documentation & Final Write-up