

# **DATA-EFFICIENT URDU SPEECH CORPUS CONSTRUCTION**

**MANAL ASGHAR & (FA24-MSDS-0004)**



**COMPUTER SCIENCE**

**FACULTY OF COMPUTING**

**MOHAMMAD ALI JINNAH UNIVERSITY**

**FALL 2025**

# DATA-EFFICIENT URDU SPEECH CORPUS CONSTRUCTION

SUBMITTED BY  
MANAL ASGHAR & (F24-MSDS-0004)

SUPERVISED BY  
DR. TAFSEER AHMED

THESIS SUBMITTED  
OF THE FACULTY  
FACULTY OF  
MOHAMMAD ALI  
IN PARTIAL  
THE REQUIREMENTS  
OF [TYPE NAME OF  
IN FULL FORM (for example MASTER OF SCIENCE IN COMPUTER  
SCIENCE)]



TO THE [TYPE NAME  
(for example  
COMPUTING)],  
JINNAH UNIVERSITY,  
FULFILMENT OF  
FOR THE DEGREE  
THE DEGREE HERE

FALL 2025

# CERTIFICATE OF APPROVAL



*It is certified that the research work presented in this thesis, titled Data-efficient Urdu speech corpus construction was conducted by Manal Asghar under the supervision of Dr. Tafseer Ahmed].*

*No part of this thesis has been submitted anywhere else for any other degree.*

*This thesis is submitted to the Department of Computer Science in partial fulfilment of the requirements for the degree of*

*Masters of Science in Fall 2025  
at the  
Mohammad Ali Jinnah University  
Karachi, Pakistan*

*[Place Date of Thesis Defense here (for example Month DD, YYYY)]*

Name of Candidate: \_\_\_\_\_

Signature: \_\_\_\_\_

## Examination Committee Members:

a) Name of External Examiner: \_\_\_\_\_

Signature: \_\_\_\_\_

\_\_\_\_\_  
Designation, Affiliation:

b) Name of Internal Examiner: \_\_\_\_\_

Signature: \_\_\_\_\_

\_\_\_\_\_  
Designation, Affiliation:

c) Name of Supervisor: \_\_\_\_\_

Signature: \_\_\_\_\_

\_\_\_\_\_  
Designation, Affiliation:

\_\_\_\_\_  
Name :

Dean, (Name of the Faculty) [for example (Faculty of  
Mohammad Ali Jinnah University, Karachi

Computing)]

# CERTIFICATE OF SUPERVISION



*This is to certify that the thesis titled, “Data-efficient Urdu speech corpus construction”, is submitted to the Department of Computer Science, Fall 2025 by Manal Asghar for the award of the degree of Master of Science in the discipline of Computer Science Data-science AI. The thesis has been carried out under my supervision. I certify that the work submitted is original and not plagiarized from any other source, except as specified in the references. Neither the thesis nor the work contained therein has been previously submitted to any other institution for a degree.*

---

**Dr. Tafseer Ahmed**

**MOHAMMAD ALI JINNAH UNIVERSITY**  
**ORIGINAL LITERARY WORK DECLARATION**

<b>Faculty</b>	<i>Faculty of Computing</i>	
<b>Program</b>	<i>MS Data Science</i>	
Student Name: Manal Asghar		Reg. No: FA24-MSDS-0004
Email: <a href="mailto:FA24MSDS0004@maju.edu.pk">FA24MSDS0004@maju.edu.pk</a>		Mobile No:0316-8036790

Research Title: Data-efficient Urdu speech corpus construction

I do solemnly and sincerely declare that:

1. I am the author of this work.
2. This work is original.
3. Use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work; I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work.
4. I hereby assign all and every rights in the copyright to this work to Mohammad Ali Jinnah University (MAJU), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of MAJU having been first had and obtained.
5. I am fully aware that if in the course of making this work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by MAJU.

Student's Signature: \_\_\_\_\_

Place: Karachi-Pakistan.

## **Acknowledgements**

I would like to express my sincere gratitude to my supervisor, Dr. Tafseer Ahmed, for his unwavering support, guidance, and encouragement throughout this research. His invaluable insights and patience have been instrumental in shaping this thesis.

I also appreciate Mohammad Ali Jinnah University and the Department of Computer Science for providing the necessary resources and a conducive learning environment that enabled me to complete this study.

A heartfelt thanks to my friends and classmates for their motivation, insightful discussions, and continuous support throughout this journey. Their encouragement has been a source of strength for me.

Finally, I am profoundly grateful to my family for their unconditional love, patience, and constant encouragement. Their support has been my greatest source of motivation.

This thesis would not have been possible without the contributions and support of these incredible individuals.

Thank you!

## **Abstract**

Text-to-Speech (TTS) and Automatic Speech Recognition (ASR) systems in low-resource languages like Urdu have continued to be a major challenge in the modern speech technology research. Whereas some languages such as English have a huge, high-quality annotated speech corpus, Urdu remains extremely resource-starved. Even though much of the Urdu text exists within the news, literature, and electronic media the absence of parallel and annotated speech material constrains scalable advances in the development of ASR and TTS. Annotation and speech data collection are time-consuming, costly, laborious and extremely manual processes and thus most past research has used small task-specific data sets or simple data augmentation methods. This leads to a vital question as to how to choose the most informative and representative reading with limited resources. To solve this issue, the present study will suggest a smart system of selecting text subsets, which will be aimed at maximizing the representativeness of the corpus without using random or heuristic sampling. The suggested method uses greedy selection methods that are conscious of distribution to build a small text subset whose lexical and contextual distributions are closely related to the entire corpus. Representativeness is assessed on statistical scales of Type coverage, Token Probability coverage and KL Divergence as well as the unigram and bigram analyses to determine both lexical and contextual coverage. The evaluation is performed on text analysis, in which the chosen subset is directly compared with the complete corpus distributions. Through experimental treatment, it has been demonstrated that a carefully chosen small subset is more balanced and information-rich than random sampling, and thus makes it a powerful basis in the future, low-resource development of ASR and TTS data. The proposed framework is feasible, affordable and can also be applied to other South Asian languages such as Sindhi, Punjabi and

Balochi which serves the larger goal of developing comprehensive and scalable speech technologies.

*Keywords:* Urdu, Low-resource languages, Text subset selection, Corpus Representative, Greedy algorithms, KL Divergence, Lexical diversity, contextual diversity.

## Table of Contents

Acknowledgements	5
Abstract	6
List of Tables	viii
List of Figures	ix
List of Abbreviations	x
Author Note	xi
CHAPTER 1: Introduction	1
1.1. DISCUSSION ON URDU SPEECH TECHNOLOGY	1
1.1.1. RESOURCE CONSTRAINTS IN URDU SPEECH DATA	1
1.1.2 IMPORTANCE OF TEXT SELECTION PRIOR TO SPEECH RECORDING	2
1.2 DISCUSSION ON DATA-EFFICIENT CORPUS CONSTRUCTION	2
1.3 TEXT SUBSET SELECTION FOR SPEECH CORPUS DESIGN	2
1.4 DISTRIBUTION-AWARE GREEDY SELECTION AND ITS BENEFITS	3
1.5 EVALUATION OF TEXT SUBSET REPRESENTATIVENESS	3
1.6 PROBLEM STATEMENT	3
 CHAPTER: LITERATURE REVIEW .....	6
METHODOLOGY .....	9
2.3 EVALUATION METHOD	11
References .....	12
PLAGIARISM VERIFICATION (for Final Submission) .....	
PLAGIARISM REPORT .....	

## **List of Tables**

## List of Figures

Figure 1 : Abc .....	00
----------------------	----

## List of Abbreviations

TTS	Text To Speech
ASR	Automatic Speech Recognition
KL Divergence	Kullback–Leibler Divergence
LRL	Low-Resource Language

Data-efficient Urdu speech corpus construction

Manal Asghar

Mohammad Ali Jinnah University

**Author Note**

Department of [Computer Science], 22-E, Block-6, P.E.C.H.S., Lal Kothi Stop,  
Main Shahrah-e-Faisal, Karachi, 75400, Sindh, Pakistan.

Email ID: [FA24MSDS0004@maju.edu.pk](mailto:FA24MSDS0004@maju.edu.pk)

## CHAPTER 1

### 1.1 DISCUSSION ON URDU SPEECH TECHNOLOGY

Urdu is a popular language that is used by millions of first and second language speakers throughout South Asia and the global diaspora. Although the Urdu language is widely used, there is very little development of strong speech technologies like Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) systems. The main reason for this limitation is the lack of large, high-quality, and representative speech corpora.

The use of data-driven methods of learning has been critical in more recent speech systems and needs a lot of paired text and speech data. Compared to high-resource languages, Urdu does not have standardized, open, and scalable speech datasets, which are decades old. This gap has significantly slowed progress in Urdu speech processing research.

#### 1.1.1 RESOURCE CONSTRAINTS IN URDU SPEECH DATA

The other most significant problem in the Urdu speech technology is that of the scarcity of annotated speech data. The available datasets are usually small-sized, limited to a particular field, or have been collected in controlled conditions which do not represent the real weaving of linguistic diversity. The process of collecting speech data requires recruiting of the speakers, the recording setup, transcription, and quality verification processes which all incur high financial cost and human resource. The enablement of high-scale corpus construction requires these limitations, particularly in low-budget and academic research efforts.

## **1.1.2 IMPORTANCE OF TEXT SELECTION PRIOR TO SPEECH RECORDING**

A basic procedure used in the building of speech corpus is text selection. The linguistic information to be recorded in the form of speech is directly dependent on the sentences that are selected to be put in writing. The problem with inefficient use of data is due to poor choice of text in spite of the amount recorded speech. In the case of low-resource languages such as the Urdu language, an intelligent text selection is necessary. A small yet representative sample of text can be picked to save much effort in recording, and maintain linguistic diversity.

## **1.2 DISCUSSION ON DATA-EFFICIENT CORPUS CONSTRUCTION**

Data-efficient corpus construction focuses on maximizing linguistic coverage while minimizing data volume. Instead of increasing dataset size, this paradigm emphasizes selecting informative and representative samples from large unannotated text corpora.

In recent years, data-efficient approaches have gained attention in low-resource NLP and speech research. These methods aim to reduce annotation costs while maintaining performance, making them particularly suitable for languages with limited resources.

## **1.3 TEXT SUBSET SELECTION FOR SPEECH CORPUS DESIGN**

Text subset selection can be used to select a fixed number of sentences in a large corpus so that the statistical characteristics of the entire dataset can be replicated in the selected subset. This methodology is particularly important in cases where very little speech is audio recordable. Using massive Urdu text corpora and only picking out the most representative sentences, small but linguistically rich speech scripts can be put together. This approach will specifically deal with the fact that there is a lot of text data and minimal volume of speech.

### 1.4 DISTRIBUTION-AWARE GREEDY SELECTION AND ITS BENEFITS

Distribution-sensitive selection algorithms attempt to align the selected subset lexical and contextual distributions to the full corpus. These methods, unlike random sampling, are more precise in their selection, taking into account word frequency and contextual patterns.

The solution proposed by greedy algorithms is practical, involving the selection of sentences in multiple steps that optimize the representativeness in the best way. These approaches have various advantages such as the enhancement of lexical coverage, enhancement of contextual balance, and minimization of redundancy. They are efficient in computation and are applicable in large scale text corpora.

### 1.5 EVALUATION OF TEXT SUBSET REPRESENTATIVENESS

It is necessary to evaluate the quality of a chosen text subset to make sure that it is representative of the entire corpus. Intrinsic evaluation metrics are usually employed as opposed to downstream model performance. Lexical and contextual representativeness are quantified by metrics like type coverage, token probability coverage, KL divergence and n-gram analysis. Through these evaluation strategies, the methods of greedy selections and the random sampling can be compared objectively.

### 1.6 PROBLEM STATEMENT

Although there is a large corpora of Urdu text, there is no known and standardized, data-efficient, and distribution-factors-aware model of a representative selection of text subsections to build a speech corpus. The current methods are either based on random selection or manual heuristics which make use of the scarce resources of recording inefficient. The study works on the issue of greedy and distribution-sensitive algorithm selection of a small but highly representative text

subset in Urdu language, the goal of which is to facilitate the scalable development of rich speech corpus to support low-resource ASR and TTS systems.

### 1.7 RESEARCH AIM AND OBJECTIVES

The main objective of this study is to create a data-effective text subset selection model in terms of the constructions of the Urdu speech corpus with limited resources. In order to accomplish this goal, the next objectives are defined:

1. In order to examine the distributional properties of large-scale Urdu text corpora.
2. To create greedy, distribution-sensitive algorithms in selecting a small but representative a subset of text.
3. To compare the representativeness of the subsets of the selected words based on both lexical and contextual similarity.
4. To determine the coverage and distributional alignment of greedy selection strategies and random sampling.
5. To create a testable methodology that can be used in future ASR and TTS data collection of Urdu.

### 1.8 SIGNIFICANCE OF THE STUDY

The study has provided contributions to the area of low-resource speech technology by moving towards thinking about the quality and representativeness of the corpus size rather than its size. The study has a high-level problem, namely the efficient data selection, as opposed to suggesting new acoustic or neural architectures. The suggested framework can be used to build small, but informative subsets of the text which can be very close to the statistics of large corpora. This will minimize the cost of recording speech directly and improve the linguistic balance of the resulting data. The approach is quite useful in both academic and industrial environments, where resources

are limited. Moreover, even though the study focuses on Urdu, the suggested solution is language-neutral and may be applied to other low-resource South Asian languages like Sindhi, Balochi, and Pashto. In this regard, the study has wider scope of implications on inclusive language technology development.

## CHAPTER 2

### 2.1 LITERATURE REVIEW

The high-quality and representative speech datasets have become a critical factor especially in the low-resource languages due to the rapid development of speech technologies. Several studies have explored strategies to bridge the data gap by leveraging text-to-speech (TTS) systems, data selection mechanisms, and efficient corpus design techniques. Yang et al. proposed the use of versatile TTS systems to enhance low-resource ASR by generating synthetic speech data, demonstrating that carefully selected synthetic data can significantly improve recognition performance when real speech data is scarce [1]. Likewise, Seki et al. also proposed a data selection framework of TTS called evaluation-in-the-loop whereby informative text samples are repeatedly sampled using model feedback, leading to greater efficiency in data and synthesis quality [2].

Recent research has also focused on dataset distillation and automated dataset construction. The DiLM framework distills large datasets into compact representations suitable for language modeling, enabling efficient learning from limited data while preserving distributional characteristics [3]. In parallel, automated end-to-end pipelines for TTS dataset generation have been proposed, integrating text processing, speaker management, and quality control to produce scalable and high-quality datasets with minimal manual intervention [4]. These approaches emphasize automation and efficiency but largely focus on high-resource or multilingual settings.

Data-efficient subset selection has been extensively studied in ASR adaptation scenarios. DITTO introduced a targeted subset selection method that balances data efficiency and fairness for accent adaptation in ASR systems, demonstrating that carefully curated subsets can outperform randomly sampled larger datasets [5]. The development of conversational speech

data with the help of TTS and large language models was also investigated in other works and emphasized the significance of text diversity and the richness of the context in which the synthetic data are created to help the speech recognition [6]. The development of Urdu-specific corpus has been relatively low. The initial attempts were oriented to the design of phonetically dense Urdu speech corpora based on manual and rule-based sentence selection methods [7][18]. The initial research on Urdu corpus construction by Becker focused on the linguistic issues of morphology and variation in scripts [15]. Newer research evaluated Urdu ASR systems and reported the constraints of the lack of sufficient and balanced training data [16]. Despite the heuristic or manual methods of selection, these works had a significant groundwork but were based on the heuristic or manual selection.

The data selection techniques of speech synthesis have also been studied on a bigger scale. Lee and Cooper contrasted speaker and utterance-based data selection approaches to TTS and found that utterance-level selection is more comprehensive and less sampling is required [8]. Gallegos et al. suggested an unsupervised method of choosing representative speakers with large multi-speaker TTS datasets to decrease redundancy and preserve the quality of synthesis [9]. Taubert et al. compared the text selection algorithms to sequence-to-sequence neural TTS and demonstrated that distribution-based selection is always better than random sampling [10]. The speech data selection research has also been affected by active learning and core-set selection techniques. Sener and Savarese proposed a core-set method of active learning which chooses representative samples using geometrical covering in feature space, which has powerful theoretical guarantees [11]. These concepts have been applied to the speech and text data selection process, but most research has been done on the vision or high-resource NLP problems.

## DATA-EFFICIENT URDU SPEECH CORPUS CONSTRUCTION

The More recent work has focused multilingual and low-resource TTS systems, including the development of advanced Urdu TTS models and surveys on efficient speech data selection [12][17]. Although these studies do not ignore the significance of representative data selection, they do not necessarily have a formal distribution-matching goal or are based on downstream performance instead of intrinsic corpus analysis.

Research Paper Name	Key Findings	Methodology	Identified Gaps
Diversity-Based Core-Set Selection for TTS (Seki et al.)	Improved TTS performance using diverse subsets	Core-set selection using linguistic and acoustic features	Requires acoustic features; not applicable before speech recording
Speech Data Selection for Efficient ASR Fine-Tuning (Rangappa et al.)	Efficient fine-tuning with fewer labeled samples	Domain classifier and pseudo-label filtering	Depends on labeled or pseudo-labeled speech data
Unsupervised Active Learning for ASR (Zheng et al.)	Reduced labeling cost through unsupervised selection	Active learning with confidence measures	Selection tied to ASR model training stage
DEFT-UCS (2024)	Data-efficient fine-tuning of language models	Unsupervised core-set selection	Focused on text editing, not speech corpus design
Developing High-Quality TTS for Punjabi and Urdu	Improved TTS quality using MMS models	Multilingual transfer learning	Relies on existing large-scale speech resources
Efficient ASR for Low-Resource Languages (Anonymous, 2025)	Performance gains via optimized training	Model-centric optimization	Does not address corpus construction stage

Table 01: Gap Analysis

## 2.2 PROPOSED METHODOLOGY

### 2.2.1 PROBLEM FORMULATION

Let  $C$  be a large text corpus and  $n$  be the desired subset size. The objective is to select a subset  $S \subset C$  such that the statistical distribution of the selected subset closely matches the distribution of the full corpus. The similarity between the two distributions is measured using distribution-based evaluation metrics.

#### 2.2.2 GREEDY KL-DIVERGENCE MINIMIZATION ALGORITHM

This method selects sentences iteratively by minimizing the KL-divergence between the full corpus distribution and the subset distribution.

**Algorithm 1:** Greedy KL-Divergence–Based Core-Set Selection from Corpus  $C$

selects a subset  $S$  of size  $n$  from the full corpus  $C$  by iteratively minimizing the KL divergence between the word distribution of the selected subset and that of the full corpus.

1.  $S \leftarrow \emptyset$
2. Compute  $P_{\text{full}}$  from  $C$
3. while  $|S| < n$  do
4.  $s^* \leftarrow \operatorname{argmin}_{\{s \in C \setminus S\}} D_{\text{KL}}(P_{\text{full}} \parallel P_{\{S \cup \{s\}\}})$
5.  $S \leftarrow S \cup \{s^*\}$
6. end while
7. return  $S$

This algorithm ensures strong theoretical alignment to the distribution being matched at the cost of being more expensive to calculate because it requires repeated corpus-wide evaluation

### 2.2.3 GREEDY DEFICIT-BASED SELECTION ALGORITHM

This method, word-level deficits between the complete corpus and that being studied are monitored and sentences that minimize such deficits prioritized.

**Algorithm 2:** Greedy Deficit-Based Core-Set Selection from Corpus C

constructs a subset  $S$  from the full corpus  $C$  by prioritizing sentences that contain under-represented words based on their distributional deficit scores.

1.  $S \leftarrow \emptyset$
2. Compute  $P_{\text{full}}$  from  $C$
3. Initialize  $\delta(w) \leftarrow P_{\text{full}}(w)$  for all words  $w$
4. while  $|S| < n$  do
5.  $s^* \leftarrow \operatorname{argmax}_{\{s \in C \setminus S\}} \sum_{\{w \in s\}} \delta(w)$
6.  $S \leftarrow S \cup \{s^*\}$
7. Update  $\delta(w)$  for  $w \in s^*$
8. end while
9. return  $S$

This is more computationally efficient and the approach intuitively represents under-represented linguistic units.

## **2.3 EVALUATION METHOD**

### **2.3.1 TYPE COVERAGE**

Measures the proportion of unique vocabulary preserved in the selected subset relative to the full corpus.

### **2.3.2 TOKEN PROBABILITY COVERAGE**

Evaluates how much probability mass of the full corpus is captured by the selected subset.

### **2.3.3 KL-DIVERGENCE**

Quantifies distributional similarity between the full corpus and the selected subset. Lower values indicate better representativeness.

### **2.3.4 N-GRAM BASED EVALUATION**

All metrics are extended to unigram and bigram levels to assess both lexical and contextual coverage.

## References

1. Y. Yang *et al.*, “Enhancing low-resource ASR through versatile TTS: Bridging the data gap,” in *Proc. International Conference on Speech Technology*, 2024.
2. K. Seki, S. Takamichi, T. Saeki, and H. Saruwatari, “Text-to-speech synthesis from dark data with evaluation-in-the-loop data selection,” in *Proc. Interspeech*, 2023.
3. A. Maekawa, S. Kosugi, K. Funakoshi, and M. Okumura,, “DiLM: Distilling dataset into language model for text-level dataset distillation,” *arXiv preprint*, 2024.
4. A. Gunduz, K. A. Yuksel, K. Darwish, G. Javadi, F. Minazzi, N. Sobieski, and S. Bratière, “An automated end-to-end open-source software for high-quality text-to-speech dataset generation,” *arXiv preprint*, 2024.
5. S. Kothawade, A. Mekala, C. S. D., M. Kothiyari, R. Iyer, G. Ramakrishnan, and P. Jyothi,, “DITTO: Data-efficient and fair targeted subset selection for ASR accent adaptation,” in *Proc. International Conference on Speech Processing*, 2023.
6. Samuele Cornell, Jordan Darefsky, Zhiyao Duan, Shinji Watanabe<sup>1</sup>, “Generating data with text-to-speech and large-language models for conversational speech recognition,” in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023.
7. A. A. Raza *et al.*, “Design and development of phonetically rich Urdu speech corpus,” in *Proc. International Conference on Language Resources and Evaluation (LREC)*, 2009.
8. Kai-Zhan. Lee and Erica. Cooper, Julia Hirschberg “A comparison of speaker-based and utterance-based data selection for text-to-speech synthesis,” in *Proc. Interspeech*, 2018.
9. P. O. Gallegos *et al.*, “An unsupervised method to select a speaker subset from large multi-speaker speech synthesis datasets,” in *Proc. Interspeech*, 2020.
10. S. Taubert *et al.*, “A comparison of text selection algorithms for sequence-to-sequence neural TTS,” in *Proc. ICSPCC*, 2022.
11. O. Sener and S. Savarese, “Active learning for convolutional neural networks: A core-set approach,” *arXiv preprint arXiv:1708.00489*, 2017.
12. S. Khan *et al.*, “Overcoming linguistic barriers: Developing advanced Urdu text-to-speech systems,” in *Proc. International Conference on Speech Technology*, 2024.
13. Srihari Bandarupalli, Bhavana Akkiraju, Charan Devarakonda, Vamsiraghushimha Narsinga<sup>1</sup> , Anil Kumar Vuppala, “Efficient ASR for low-resource languages,” *arXiv preprint*, 2025.

14. K. Seki, S. Takamichi, T. Saeki, and H. Saruwatari, “Active learning for text-to-speech synthesis with informative sample collection,” in *Proc. Interspeech*, 2021.
15. D. Becker, *A Study in Urdu Corpus Construction*. Berlin, Germany: Academic Press, 2002.
16. S. Arif, A. J. Khan, M. Abbas, A. A. Raza, and A. Athar, “WER We Stand: Benchmarking Urdu ASR Models,” in *Proc. 31st Int. Conf. Computational Linguistics (COLING)*, Abu Dhabi, UAE, 2025, pp. 5952–5961.
17. A. H. Azeemi, I. A. Qazi, and A. A. Raza, “A survey on data selection for efficient speech processing,” *IEEE Access*, 2025.
18. K. Seki *et al.*, “Diversity-based core-set selection for text-to-speech with linguistic and acoustic features,” in *Proc. Interspeech*, 2022.
19. S. Rangappa *et al.*, “Speech data selection for efficient ASR fine-tuning using domain classifier and pseudo-label filtering,” in *Proc. IEEE ASRU*, 2023.
20. Y. Zheng *et al.*, “Unsupervised active learning: Optimizing labeling cost-effectiveness for automatic speech recognition,” in *Proc. IEEE ICASSP*, 2024.
21. D. Das and V. Khetan, “DEFT-UCS: Data efficient fine-tuning for pre-trained language models via unsupervised core-set selection for text-editing,” *arXiv preprint*, 2024.
22. F. Naseem, M. Sajid, F. Adeeba, S. Rauf, A. Mustafa, and S. Hussain, “Developing high-quality TTS for Punjabi and Urdu against MMS model,” Technical Report (PDF), 2024.
23. K. Seki, S. Takamichi, T. Saeki, and H. Saruwatari, “TTSOps: A closed-loop corpus optimization framework for training multi-speaker TTS models from dark data,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.
24. S. Hussain, P. Neekhara, X. Yang, E. Casanova, S. Ghosh, R. Fejgin, R. Langman, M. Desta, L. Tavabi, and J. Li, “ALIGN2SPEAK: Improving TTS for low resource languages via ASR-guided online preference optimization,” *arXiv preprint arXiv:2509.21718v1 [cs]*, 2025.

### Footnotes

<sup>1</sup>[Add footnotes, if any, on their own page following references. For APA formatting requirements, it's easy to just type your own footnote references and notes. To format a footnote reference, select the number and then, on the Home tab, in the Styles gallery, click Footnote Reference. The body of a footnote, such as this example, uses the Normal text style. *(Note: If you delete this sample footnote, don't forget to delete its in-text reference as well. That's at the end of the sample Heading 2 paragraph on the first page of body content in this template.)*]

Table 1

[Table Title]

Column Head	Column Head	Column Head	Column Head	Column Head
Row Head	123	123	123	123
Row Head	456	456	456	456
Row Head	789	789	789	789
Row Head	123	123	123	123
Row Head	456	456	456	456
Row Head	789	789	789	789

*Note:* [Place all tables in their own section, following references (and, if applicable, footnotes).

Start a new page for each table, include a table number and table title for each, as shown on this page. All explanatory text appears in a table note that follows the table, such as this one. Use the Table/Figure style, available on the Home tab, in the Styles gallery, to get the spacing between table and note. Tables in APA format can use single or 1.5 line spacing. Include a heading for every row and column, even if the content seems obvious. A default table style has been setup for this template that fits APA guidelines. To insert a table, on the Insert tab, click Table.]

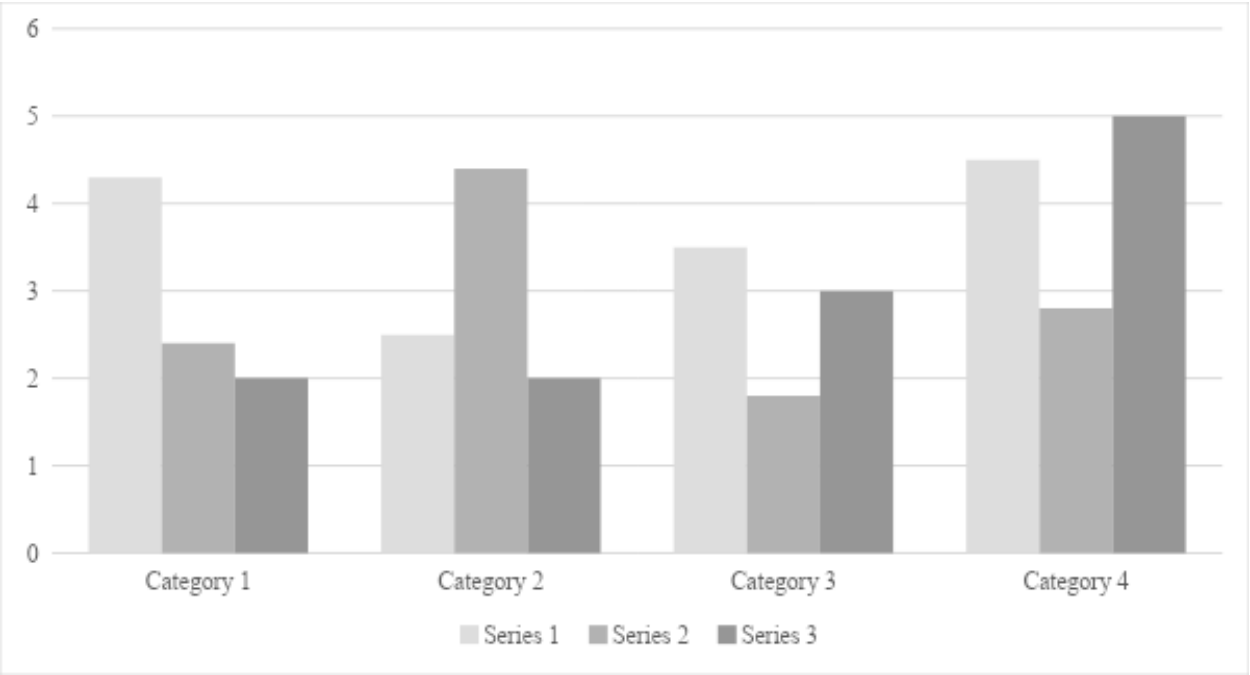


Figure 1. [Figure Title]

[Include all figures in their own section, following references (and footnotes and tables, if applicable). Include a numbered caption for each figure. Use the Table/Figure style for easy spacing between figure and caption.]

## **Appendix**

For any supplementary material that supervisor considers important and should be included. Examples include FORMS, CODING etc.

## PLAGIARISM VERIFICATION (for Final Submission)

(Note: This form is filled at the final submission of Thesis/Research Project Reports for graduation.)

Title of Research Thesis: \_\_\_\_\_

Total Pages: \_\_\_\_\_

Student Name : \_\_\_\_\_

Registration ID : \_\_\_\_\_

Supervisor: \_\_\_\_\_  
\_\_\_\_\_

Program: \_\_\_\_\_

Faculty: \_\_\_\_\_

This is to report that the above document was scanned for similarity detection and has been submitted in turnitin repository. Process and outcome is given below:

Software used: **Turnitin** \_\_\_\_\_

Date: \_\_\_\_\_

Similarity Index: \_\_\_\_\_

Total word count: \_\_\_\_\_

File Name: \_\_\_\_\_

Digital Receipt No.: \_\_\_\_\_

\_\_\_\_\_  
Signature and Stamp of the  
Office of Research & Project

## PLAGIARISM REPORT

The Plagiarism Report issued shall be attached here