Language Agnostic Phone-Based Audio Search

Abhishek Chippa, Vishnu Alla, Yashwanth Katamreddy, Daniel Whitenack, Matthew A. Lanham

° Purdue University, Krannert School of Management; † SIL International

°achippa@purdue.edu; ° valla@purdue.edu; ° ykatamre@purdue.edu; † dan_whitenack@sil.org; ° lanhamm@purdue.edu

ABSTRACT

When checking the quality of content translated into a set of audio recordings, translators or translation consultants need to search for keywords contained in the audio recordings. Currently, these individuals are required to manually scour through hours of recordings to find those containing keywords that need to be modified. This project aims to implement phone-based audio search to automate the process of finding the keywords and alleviate the burden on translators. By performing this search in phone space using universal phone recognizer, the audio search can be applied to content translated to any language. *(A phone is a distinct speech sound in a language).*

INTRODUCTION

- The goal of this project is to **automate the process of finding keywords in audio recordings regardless of language**. Such an audio search will increase the pace of oral translation quality assessment. (Fig.1 shows conservative estimates of the times needed to review and search for keywords in various audio books)
- **Language agnostic phone-based audio search** is used instead of Speech recognition models to extend the functionality to languages without automatic speech recognition support.

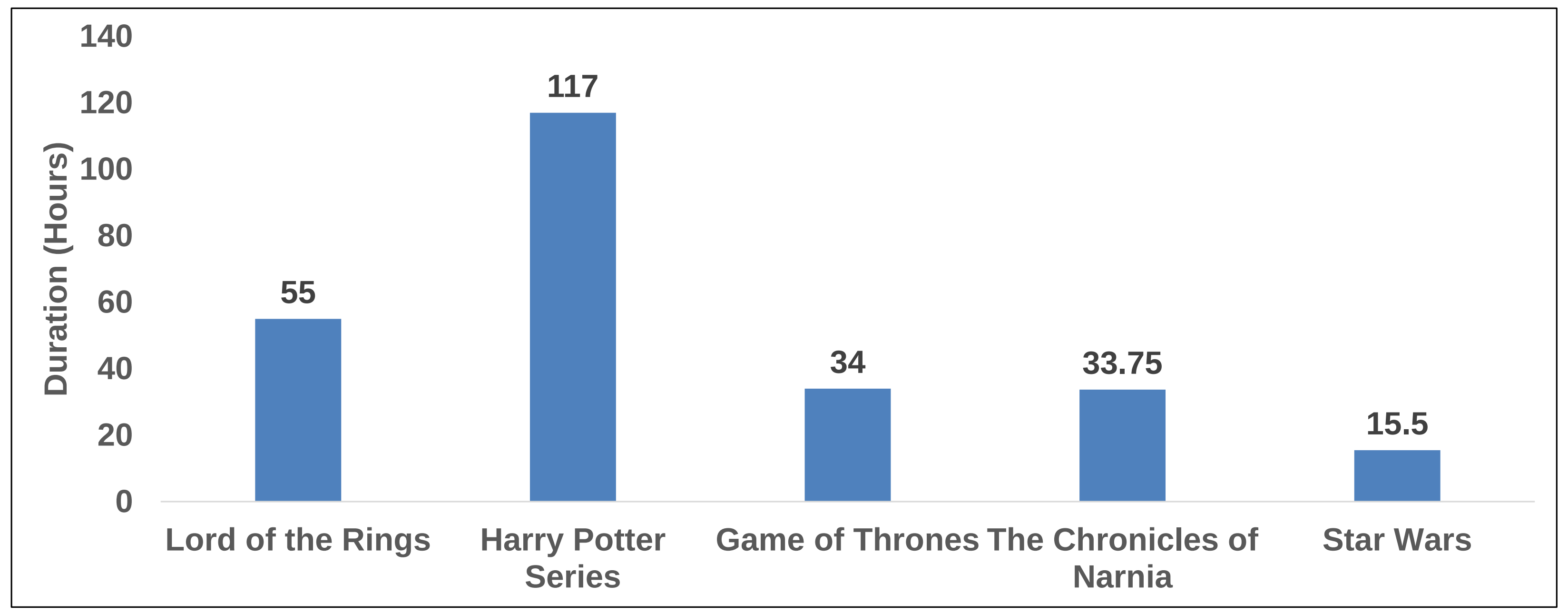


Fig 1. Time required to search through audio books

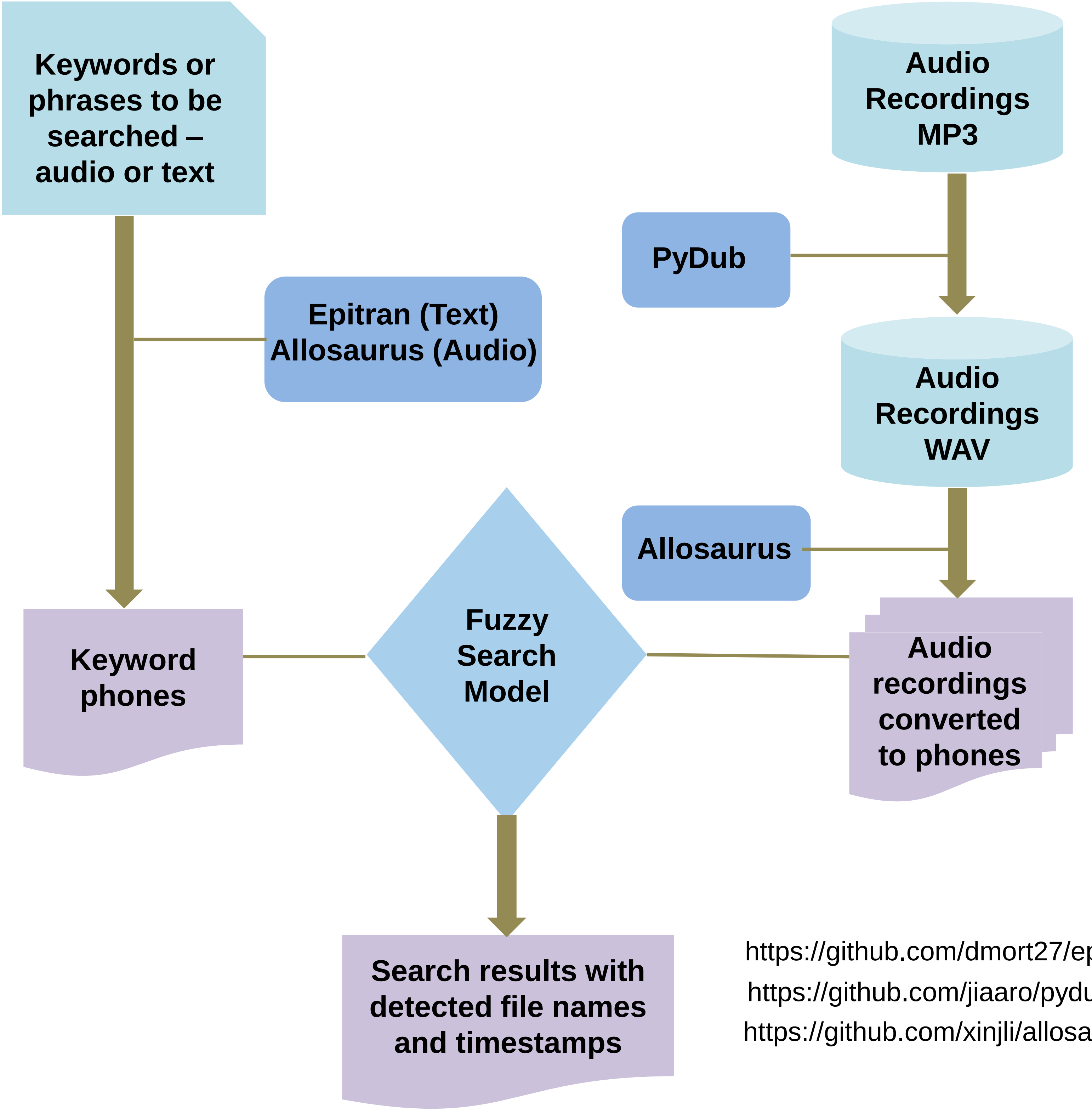
Research Question

- Can we search through audio recordings in phone space rather than in text space? If so, can we provide searching functionality for local language recordings?

LITERATURE REVIEW

Author	Summary	Methodology
X Li et al. (2020)	Universal Phone Recognition with a Multilingual Allophone System	The paper implemented 3 models: Shared Phone, Private Phone and Allosaurus, and compared error rates across the models. Suggested Allosaurus model had 17% higher accuracy.
K Siminyu et al. (2021)	Phoneme Recognition through Fine-Tuning of Phonetic Representations	This paper fine-tuned the pre-trained Allosaurus model by working on Bukusu and Saamia languages achieving accuracy increased by 90% and 75% respectively. It observed that fine-tuning of Allosaurus model is promising with as few as 100 utterances.

METHODOLOGY


Keywords or phrases to be searched – audio or text
Audio Recordings MP3
PyDub
Epitran (Text) Allosaurus (Audio)
Audio Recordings WAV
Allosaurus
Keyword phones
Fuzzy Search Model
Audio recordings converted to phones
Search results with detected file names and timestamps

https://github.com/dmort27/epitran
https://github.com/jiaaro/pydub
https://github.com/xinjli/allosaurus

Fig 2. Process pipeline for phone-based search

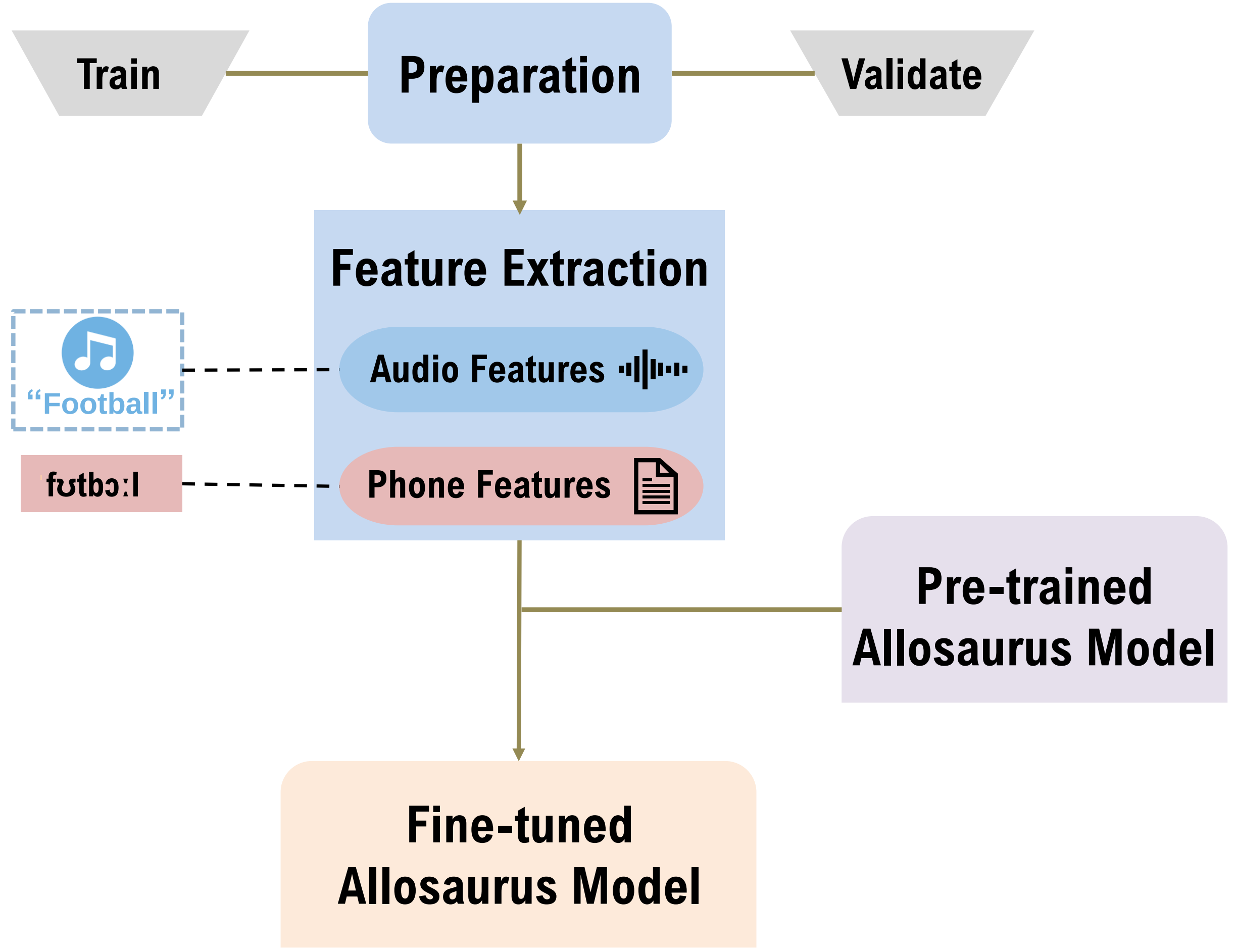
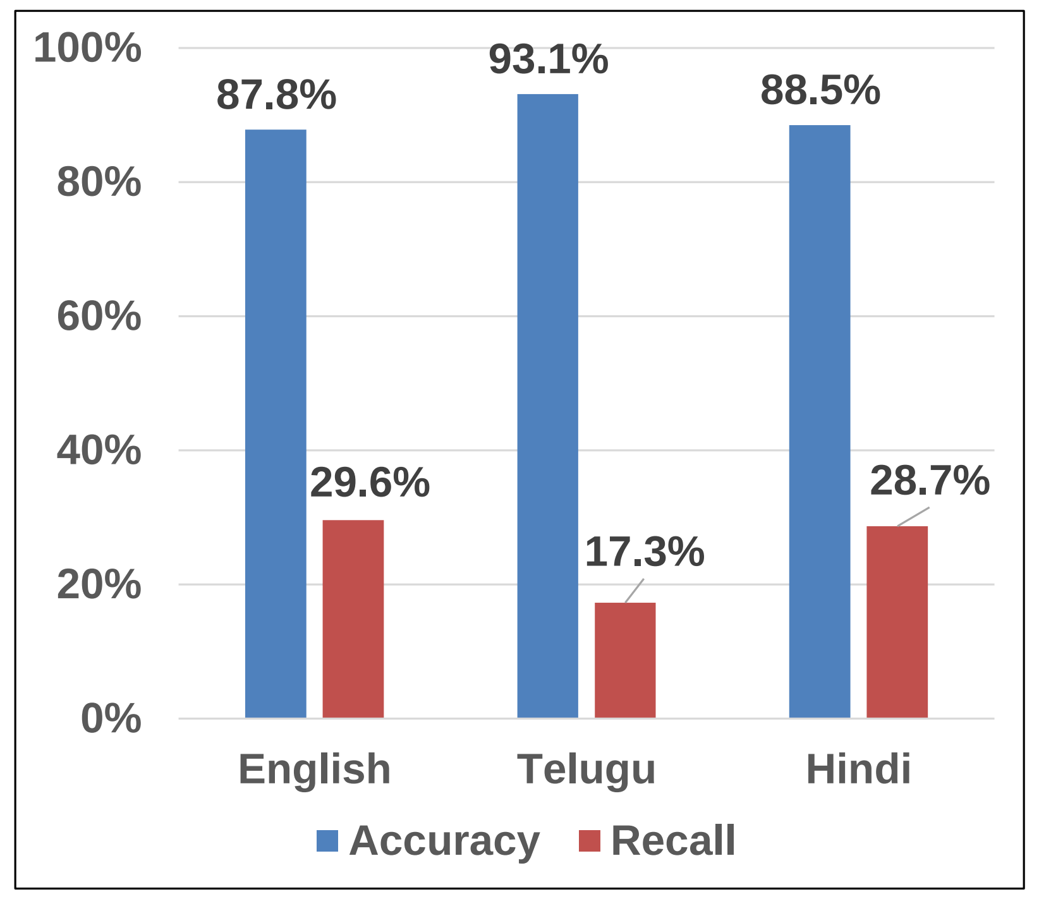


Fig 3. Allosaurus model finetuning process to improve the performance of phone recognition from audio files

RESULTS



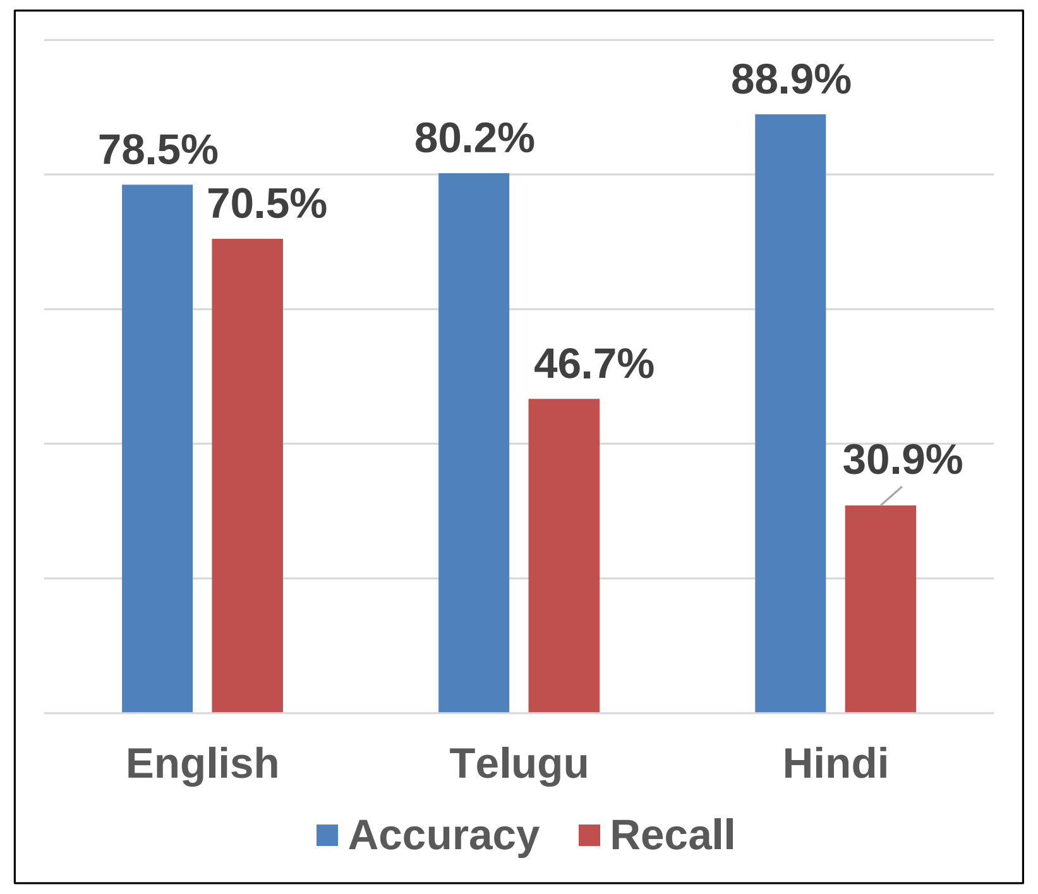


Fig 4. Model results – Universal model

Fig 5. Model results – Fine-tuned model

BROADER IMPLICATIONS

- Implementing the phone-based audio search will result in significant cost savings and help to serve communities that speak languages without speech recognition support.

Book	Billing Rate ($ per hr)	Duration (Hours)	Cost Savings ($)
Lord of the Rings	$75	55	$4,125
Harry Potter Series	$75	117	$8,775
Game of Thrones	$75	34	$2,550
The Chronicles of Narnia	$75	34	$2,531
Star Wars	$75	16	$1,163

Fig 6. Cost Savings ($)

KEY TAKEWAYS

- After fine-tuning the Allosaurus phone recognition model, we noticed a significant increase in the recall percentages. Therefore, Phone-based audio search is a promising approach to search through audio files in languages without speech recognition support.
- In future, the model can be deployed as a REST API for practical use.

ACKNOWLEDGEMENTS

We would like to thank Professors Matthew Lanham and Daniel Whitenack for their guidance and constant support on this project.