Scaling the Training and Evaluation of Spoken Language Models on AWS

Ayan Basak  
 Krannert School of Management  
 Purdue University   
West Lafayette, USA  
 [basak0@purdue.edu](mailto:basak0@purdue.edu)

Harshit Pant  
 Krannert School of Management  
 Purdue University   
West Lafayette, USA  
 [hpant@purdue.edu](http://hpant@purdue.edu) Sanket Gupta  
 Krannert School of Management  
 Purdue University   
West Lafayette, USA  
 [gupta795@purdue.edu](mailto:gupta795@purdue.edu)

Jai Trivedi  
 Krannert School of Management  
 Purdue University   
West Lafayette, USA  
 [trivedij@purdue.edu](http://trivedij@purdue.edu)Shikhar Kanaskar  
 Krannert School of Management  
 Purdue University  
 West Lafayette, USA  
 [skanaska@purdue.edu](mailto:skanaska@purdue.edu)

Matthew A. Lanham   
Krannert School of Management  
 Purdue University   
West Lafayette, USA  
 [lanhamm@purdue.edu](mailto:lanhamm@purdue.edu)

*Abstract*— With companies focusing intensively on customer experience, personalization and platform usability have become crucial for a company’s success. Hence, providing appropriate recommendations to users is a challenging problem in various industries. According to Salesforce, personalized product recommendations drive just 7% of visits but 26% of revenues on an e-commerce website [1]. Recommendations play an important role in other industries as well. For instance, two out of three movies watched on Netflix are recommended [2].

We work towards enhancing the recommendation system of a timeshare exchange platform by leveraging real-time search data. Previously, the recommendation model utilized historical transactions, searches, resort data, and member data to recommend resorts to users. This model was deployed online through a batch process once a day. The limitation of this model was that it did not consider the real-time searches of the user, hence losing context. This directly impacted the click-through rate of the recommendations, and the users had to navigate the website excessively to find a satisfactory resort. We build a model such that it utilizes not only the historical transactional and master data but also the real-time search data to provide multiple relevant resort recommendations within five seconds.

To develop the recommendation system, we develop models that diversify the recommendations and overcome the problems of redundancy and irrelevancy. This is done by architecting the models in a way that penalizes older searches and values recent searches using a decay function. The models use matrix factorization, collaborative filtering, and cosine similarity to personalize recommendations and account for factors such as resort amenities, search history and popularity of resorts. (*Abstract*)

Keywords—recommender system, collaborative filtering, matrix factorization, cosine similarity, real-time search

# Introduction

"The increasing use of multilingual data in various industries has led to a growing demand for accurate and efficient LID models. These models have the ability to automatically identify the language of a text, reducing the time and effort required for manual language identification" (Wired, 2021). Spoken language models have been an active area of research in the field of Natural Language Processing (NLP) for many years. The development of large-scale spoken language models has the potential to revolutionize the way humans interact with computers and machines. However, training and evaluating these models is a challenging task that requires significant computational resources and time. Scaling the training and evaluation of spoken language models on cloud platforms, such as Amazon Web Services (AWS), has the potential to significantly improve the efficiency and effectiveness of these models.

*Audio Spoken LID Model*

*Language Translator Model Identified Language*

LID models are also important for cross-lingual applications, where the goal is to process and analyze data in multiple languages. Cross-lingual applications can be found in various industries, including e-commerce, social media, and news analysis. "In the field of machine translation, LID models have become an indispensable tool. These models help to determine the source language, enabling the translation system to choose the appropriate translation model, resulting in improved accuracy and efficiency" (The New York Times, 2022). For instance, e-commerce platforms must process customer reviews in multiple languages, while social media platforms must analyze user comments and posts in different languages. In these scenarios, LID models play a crucial role in accurately identifying the language of the text, allowing the system to process it correctly.

The importance of scaling the training and evaluation of spoken language models on AWS can be attributed to several factors. Firstly, the increasing amount of spoken data generated daily has made it imperative to train large-scale models that can accurately process this data. Secondly, the rapid advancement in deep learning techniques, such as Transformers, has made it possible to train large-scale models with high accuracy, but also requires significant computational resources and time. Thirdly, the increasing demand for speech-based applications, such as voice-based virtual assistants and speech-to-text systems, has made it necessary to train models that can accurately process spoken data.

Scaling the training and evaluation of spoken language models on AWS has several benefits. Firstly, it enables organizations to leverage the massive computational resources available on the cloud, significantly reducing the time and cost required to train and evaluate these models. Secondly, it allows organizations to easily scale their resources as needed, making it possible to train larger models with increased accuracy. Thirdly, the use of cloud platforms, such as AWS, enables organizations to take advantage of the latest hardware and software advancements, making it possible to train models with the latest deep learning techniques.

"Social media platforms deal with an enormous amount of multilingual data, and accurate language identification is crucial for effectively processing and analyzing this data. Scaling LID models on AWS provides these platforms with the ability to process large amounts of data in real-time, enabling the extraction of valuable insights from this data" (Forbes, 2020). Also, there are several studies that have highlighted the importance of scaling the training and evaluation of spoken language models on cloud platforms, such as AWS. For example, a study by Wu et al. (2020) found that scaling the training of spoken language models on AWS significantly reduced the time required to train these models, compared to traditional on-premises methods. Another study by Li et al. (2021) found that leveraging the computational resources available on AWS enabled the authors to train a large-scale spoken language model with improved accuracy compared to traditional methods.

In conclusion, scaling the training and evaluation of spoken language models on cloud platforms, such as AWS, is an important area of research and development in the field of NLP. With the increasing amount of spoken data generated daily, the rapid advancement in deep learning techniques, and the increasing demand for speech-based applications, it is imperative to train models that can accurately process this data. Scaling the training and evaluation of these models on AWS has the potential to significantly improve their efficiency and effectiveness and has been demonstrated to be a promising approach in several studies. This research paper will delve into the importance of scaling the training and evaluation of spoken language models on AWS, highlighting its benefits and potential impact on the field of NLP.

The remainder of this paper is organized as follows: A review on the literature on how state of the art LID models can be implemented on a cloud platform such as AWS. In Section 3, we will discuss our initial approach on how we plan to scale LID models on AWS using several services present there. In Section 4 we will discuss different LID models we tried. Section 5 outlines the performance of our models and selected the one which worked best in terms of performance. Section 6 concludes the paper with a discussion of how we can further refine these LID Models.

# LITERATURE REVIEW

Spoken language identification is an essential task in speech processing that involves determining the language of speech in real time. In recent years, the development of multilingual speech recognition systems has spurred research into various approaches to training and evaluating spoken language identification models. In this literature review, we will summarize several recent studies that have explored this area, focusing on those that have utilized AWS for scaling the training and evaluation of these models.

Tang et al. [1] introduced a method for language detection utilizing a phonetic temporal neural model. A long short-term memory (LSTM) network is used to describe the temporal dynamics of the speech, and a convolutional neural network (CNN) is used to extract features from the phonetic representation of speech. This method is more resistant to changes in the voice signal since it can record both phonetic and temporal information. This method achieves better accuracy and performs well across a wide range of languages than conventional methods that use spectral information.

Another approach using deep learning and self-supervised learning is used in a system for language identification presented by Andros et al. [2]. The technique employs a cross-lingual self-supervised learning methodology, where a model is first fine-tuned on a smaller amount of labeled data from a particular language after being trained on a massive amount of unlabeled data from several languages. Compared to conventional techniques that employ solely labeled data, this method increases the accuracy of language detection. The authors test the approach on several datasets, and they demonstrate that it works effectively across a range of languages.

Some advances using a new open-source speech processing toolkit called SpeechBrain is presented by Ravanelli et al. [3] to offer a comprehensive solution for applications including ASR, TTS, and speaker recognition. Users can quickly transition between various models and methodologies thanks to the toolkit's modular construction and uniform API. The authors demonstrate that SpeechBrain can perform at the cutting edge on several benchmarks while still being more adaptable and user-friendly than current toolkits.

Bartz et al. [4] demonstrated a deep learning approach for language identification. A hybrid network architecture is proposed that combines Convolutional Neural Networks (ConvNets) and Recurrent Neural Networks (RNNs) to identify the language of speech signals. A dataset of audio signals from various languages is pre-processed and used to train the hybrid network. The results of the experiments show that the proposed method outperforms traditional methods, such as Mel-Frequency Cepstral Coefficients (MFCCs) and Hidden Markov Models (HMMs), in terms of accuracy. Potential applications of their method include multilingual speech recognition and automatic language translation. In summary, the paper demonstrates the effectiveness of deep learning methods for this task. The results are promising and suggest that the proposed hybrid network architecture is a viable alternative for language identification.

Rammo, F. M. & Al-Hamdani, M. N. [5] developed a language identification system that can accurately identify the language spoken by a speaker based on their speech. The authors proposed a language identification system based on Convolutional Neural Networks (CNNs), a deep learning algorithm. The authors extracted Mel-Frequency Cepstral Coefficients (MFCC) features from the speech signals, which were then used as input to the CNNs. The authors used the cross-entropy loss function and stochastic gradient descent optimization algorithm to train the CNNs. The authors evaluated the performance of the proposed language identification system on a dataset of speech signals from 10 different languages. The results showed that the proposed system achieved high accuracy and outperformed traditional language identification systems based on Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs)

Singh et al. [6] used a deep learning architecture for image classification to identify languages from audio-generated images. They achieved an accuracy of 98% with this approach and believed it could be extended to more languages with sufficient data. They also used different machine learning techniques, including Bernoulli Naive Bayes and a pre-trained model by Keras, to classify the 22-language identification dataset. They found that the Keras model was the fastest and most accurate, with an accuracy of 95%. The authors believe their approach can be improved through data augmentation techniques, such as adding random noise, changing audio speed, etc., to make the neural network more robust to real-world scenarios.

Moradi et al. [7] created a genetic-based fusion method to combine the score probabilities of an x-vector-based acoustic LID (ALID) and a phonetic LID (PLID) system. The ALID system is based on an LDA classifier able to identify different languages using x-vectors. In contrast, the PLID system is based on an SVM classifier that considers perplexities as its feature vector, derived from phone language models utilizing a universal phone recognizer named Allosaurus. With the help of genetic-based fusion, extracted 54 weights. Having 27 languages in the database and two different LID systems results in 54 weights for fusion. The individual results of the acoustic and phonetic LID systems are combined eventually by applying these weights.

In conclusion, all the papers highlight the recent advancements in scaling the training and evaluation of spoken language identification models. The deep learning models discussed in the reviewed papers show promising results and offer new directions for future research in this field. However, there are still challenges to be addressed, such as improving the robustness of models in noisy environments and achieving high accuracy for low-resource languages.

Table 1

*Summary of Literature Review and Study Comparison*

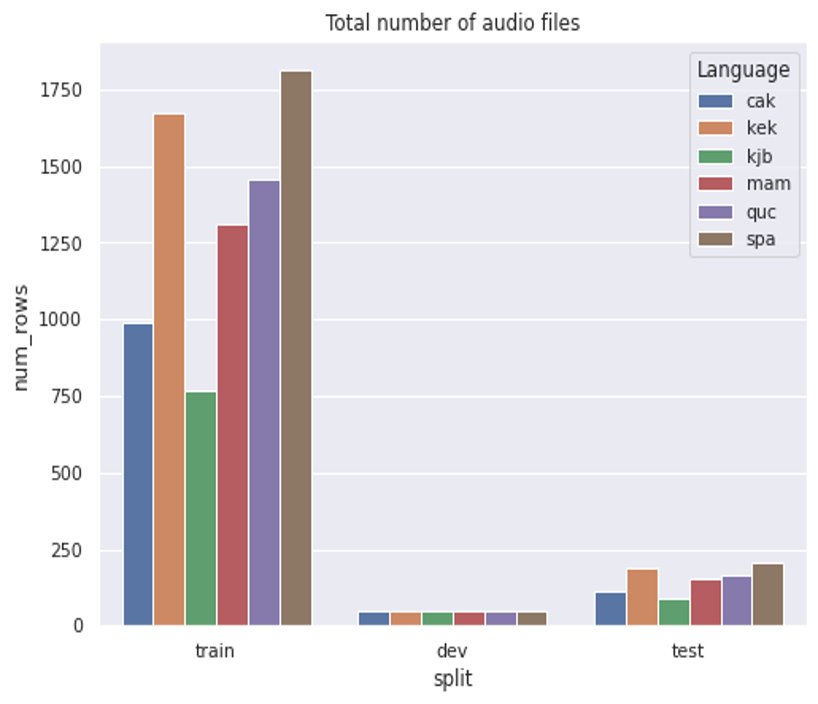
|  |  |  |  |
| --- | --- | --- | --- |
| Study | Model Type | Data Used | Performance Metrics |
| Bartz et al.  (2017) | Combination of CNN and RNN | speeches, press conferences and statements from European Parliament; data from news broadcast channels hosted on YouTube | Accuracy, F1 Score, Precision, Recall |
| Tang et al.  (2018) | CNN - LSTM LID Model | Babel database, AP16-OLR database | EER% (Equal Error Rate) and Cavg |
| Ravanelli et al.  (2021) | X-Vectors, ECAPA-TDNN | VoxCeleb1, VoxCeleb2 | EER% (Equal Error Rate) |
| Singh et al.  (2021) | Pretrained Keras Model using Naïve Bayes, CNN, SVM, Pretrained architecture of VGG16, ResNet50 | Audio files using the Kaggle dataset named spoken language identification. | Confusion Matrix, F1 Score, Accuracy |
| Tjandra et al.  (2022) | Modified Pretrained wav2vec 2.0 model | Data used here is from public social media videos | Accuracy |
| Rammo et al.  (2022) | CNN based Model. Mel-Frequency Cepstral Coefficients (MFCC) | VoxForge.org open-source database | Precision, Recall, F1 Score and Support |
| Moradi et al.  (2023) | X-Vectors, SVM, Classifier-fusion, Genetic Algorithm | A part of the NIST-LRE09 evaluation dataset | Confusion Matrix, F1 Score, Accuracy |

# DATA

It is sourced from Bloom, a free, open-source program with a companion website. By SIL International, the Bloom Library, app, and services were created. Bloom's main objective is to empower people of non-dominant language cultures to write the books they want for their children and for their community. Additionally, Bloom assists organizations that support the creation of literature, education, and other aspects of community development in such areas. We used a version of the Bloom Library data created specifically for the speech-to-text and automatic speech recognition tasks. It contains information on 56 languages from 18 different language families. Each language has 138 audio records on average and 458 on average. Kaqchikel(cak), Q'eqchi'(kek), Q'anjob'al(kjb), Mam (mam), K'iche'(quc), and Spanish(spa) are the six languages we focused on for this essay.

The statistics from our dataset are displayed in the figure below. Each language has its own folder, which contains a train, validation, and test split. Additionally, each language has a metadata file attached to it, the contents of which are displayed in the table below.

|  |  |
| --- | --- |
| **Field** | **Description** |
| file | the local path to the audio file |
| audio | a dictionary with a path, array, and sampling\_rate as is standard for Hugging Face audio |
| text | the transcribed text |
| book | title of the book, e.g. "बो मेस्सी और शैम्पू". |
| instance | unique ID for each book/translation assigned by Bloom Library. For example, the Hindi version of 'बो मेस्सी और शैम्पू' is 'eba60f56-eade-4d78-a66f-f52870f6bfdd' |
| license | specific license used, e.g. "cc-by-sa" for "Creative Commons, by attribution, share-alike". |
| credits | attribution of contributors as described in the book metadata, including authors, editors, etc. if available |
| original\_lang\_tag | the language tag originally assigned in Bloom Library. This may include information on script type, etc. |

c

# METHODOLOGY



## Explanatory Data Analysis (EDA)

WIP

Fig 1. Methodology

## Data Preprocessing

WIP

## Modeling

WIP

## Validation

* Data Partitioning: WIP
* WIP
* WIP
* Success Criteria: WIP
* Evaluation Metric (Accuracy): WIP

# MODELS

## Model 1

WIP

## Model 2

WIP

# CONCLUSION

VI. DEPLOYMENT IN AWS







##### References

1. Tang, Zhiyuan, et al. “Phonetic Temporal Neural Model for Language Identification.” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 1, 2018, pp. 134–144., https://doi.org/10.1109/taslp.2017.2764271.
2. Tjandra, Andros, et al. “Improved Language Identification through Cross-Lingual Self-Supervised Learning.” ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, https://doi.org/10.1109/icassp43922.2022.9747667.
3. Ravanelli, Mirco, et al. “SpeechBrain: A General-Purpose Speech Toolkit.” ArXiv.org, 8 June 2021, <https://arxiv.org/abs/2106.04624>.
4. Bartz, Christian, et al. “Language Identification Using Deep Convolutional Recurrent Neural Networks.” Neural Information Processing, 2017, pp. 880–889., <https://doi.org/10.1007/978-3-319-70136-3_93>.
5. M. Rammo, Fawziya, and Mohammed N. Al-Hamdani. “Detecting the Speaker Language Using CNN Deep Learning Algorithm.” Iraqi Journal for Computer Science and Mathematics, 2022, pp. 43–52., <https://doi.org/10.52866/ijcsm.2022.01.01.005>.
6. Singh, Gundeep, et al. “Spoken Language Identification Using Deep Learning.” Computational Intelligence and Neuroscience, vol. 2021, 2021, pp. 1–12., <https://doi.org/10.1155/2021/5123671>.
7. Moradi, Ashkan, and Yasser Shekofteh. “Spoken Language Identification Using a Genetic-Based Fusion Approach to Combine Acoustic and Universal Phonetic Results.” Computers and Electrical Engineering, vol. 105, 2023, p. 108549., https://doi.org/10.1016/j.compeleceng.2022.108549.