IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134212

Language identification for homophonic Short utterance using CNN

Karthikraj ghorpade[1], Anilkumar[2], Pratham Chavan[3], Kumar arayan[4]

Computer Science and Engineering Dayananda Sagar University Bengaluru, India[1-4]

Abstract: We propose a novel approach for language identification, specifically tailored for the challenging task of distinguishing homophonic short utterances. Homophonic utterances, where different languages produce similar sounds, pose a significant challenge in multilingual speech processing. We introduce a Convolutional Neural Network (CNN) architecture optimized for extracting discriminative features from audio segments. These homophonic utterances, characterized by similar sounds across different languages, are notoriously difficult to distinguish and thus require specialized techniques in multilingual speech processing.

Experimental results demonstrate the superiority of our CNN-based approach in language identification, making it a valuable contribution to the field of multilingual speech processing. The experimental study was carried out on a real-time dataset comprising Hindi, Kannada, Telugu, Marathi, and several other languages. In addition to the CNN-based approach, we also employed three traditional classifiers: Deep Learning, Convolutional Neural Network, and others. Experimental evaluations underscore the effectiveness of the CNN-based approach, showcasing its ability to achieve impressive accuracy in identifying languages within homophonic contexts. To provide a comprehensive assessment, we implemented approaches for different duration intervals, including 5 seconds, 10 seconds, and 20 seconds.

This innovative methodology addresses a pressing challenge in language identification, particularly in the context of homophonic utterances, and offers a promising solution for multilingual speech processing. Through rigorous experimentation and comparative analysis, our approach demonstrates notable advancements in accuracy and performance, thereby contributing significantly to the field of language identification and multilingual speech processing.

Keywords: Multilingual speech processing, CNN, Real-time datasets Hindi, Kannada, Telugu, Marathi, tamil ,Urdu, 5 sec, 10 sec, 20 sec, Deep learning.

I. INTRODUCTION

Language identification (LID) serves as a cornerstone in various applications such as speech recognition, voice assistants, and multilingual content processing. While conventional systems rely on handcrafted acoustic features, Convolutional Neural Networks (CNNs) offer effectiveness akin to computer vision tasks. Integration of CNN-based LID models into applications like multilingual voice assistants and customer service facilitates enhanced functionality and accuracy. The primary objective of this project is to leverage CNNs for precise language identification, particularly in scenarios involving homophonic short utterances. To harness the power of deep learning for LID, a dataset comprising homophonic short utterances across different languages is compiled and converted into a suitable format, such as spectrograms or time-frequency representations. CNNs, known for capturing spatial patterns in audio, form the crux of the model architecture. The typical CNN architecture involves multiple layers adept at discerning intricate linguistic nuances within the audio data.

The CNN model undergoes training using a labeled dataset of homophonic short utterances, during which it learns to discriminate between languages employing optimization algorithms like stochastic gradient descent. Subsequently, the model is rigorously tested on a separate validation dataset to ascertain real-world accuracy. Post-processing techniques, including language modeling and sequence- based methods, are employed to further enhance language identification accuracy, particularly for short or ambiguous utterances.

In an increasingly multilingual society, accurate language identification holds pivotal importance. The integration of CNN-based LID models into digital voice assistants and government services facilitates seamless communication and accessibility across diverse linguistic landscapes. Enhanced access to language-specific services and content, spanning domains such as education and healthcare, amplifies societal inclusivity and empowerment.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134212

The deployment of machine learning models, albeit beneficial, entails substantial energy consumption, thereby necessitating optimization for energy efficiency. Green data center practices play a vital role in curbing environmental consequences arising from data center operations. Additionally, responsible management of electronic hardware, coupled with recycling initiatives, mitigates e- waste generation. Minimizing carbon emissions through efficient algorithms and data management strategies contributes towards a sustainable ecosystem.

The challenge lies in accurately identifying languages within homophonic short utterances, where distinct languages produce similar sounds, thereby posing a formidable obstacle in multilingual speech processing. Traditional methods often falter in distinguishing between languages in such scenarios, underscoring the need for robust language identification techniques. This article addresses the burgeoning demand for precise spoken language identification in multilingual speech recognition, proposing a CNN-based solution to mitigate the aforementioned challenges and ensure high accuracy for multilingual interactions and language-specific services.

II. LITERATURE SURVEY

Some Conclusions we got from the necessary references:

1. Language identification-Bases evaluation of single channel speech seperation of overlapped Speeches (2022): This paper investigates a two-stage training scheme that combines speech separation and language identification tasks in multi-lingual environments. The separated speech is fed to a language identification engine, which evaluates its accuracy. The model used is a single-channel speech separation network trained with WSJ0-2mix. The study found that Chinese, Japanese, Korean, Indonesian, and Vietnamese speakers showed significantly improved recognition results when the language identification network model was based on single-person single-speech frequency spectrum features.

2. Utterance-level end-to-end language identification using attention-baesd CNN-BLSTM (2022):

The paper introduces an attention-based Convolutional Neural Network-Bidirectional Long- short Term Memory (CNN-BLSTM) for end-to-end language identification The model is performed at the utterance level, allowing direct decision-making. It is combined with a self-attentive pooling layer for speech utterances of arbitrary duration Experiments show comparable error reduction with other neural network approaches .

3. Spoken language Recognition:from fundamental to practice(2022):

Spoken language recognition refers to the automatic process through which we determine or verify the identity of the language spoken in a speech sample. We study a computational framework that allows such a decision to be made in a quantitative manner. In recent decades, we have made tremendous progress in spoken language recognition, which benefited from technological breakthroughs in related areas, such as signal Processing, pattern recognition, cognitive science, and machine learning. In this paper, we attempt to provide an introductory tutorial on the fundamentals of the theory and the state-of-the-art solutions, from both phonological and computational aspects. We also give a comprehensive review of current trends and future research directions using the language recognition evaluation (LRE) formulated by the National Institute of Standards and Technology (NIST) as the case studies.

4. Spoken Language Identification System Using Convolutional Recurrent Neural Network (2023):

In this article, we present a novel spoken language identification system leveraging a hybrid Convolutional Recurrent Neural Network (CRNN) architecture, integrating Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) components.

The system is evaluated on seven languages, including Arabic, sourced from subsets of the Mozilla Common Voice (MCV) corpus. Through comparative analysis of Gammatone Cepstral Coefficient (GTCC) and Mel Frequency Cepstral Coefficient (MFCC) features, as well as their combination, at the feature extraction stage, our findings demonstrate superior performance when utilizing combined GTCC and MFCC features over individual implementations. With an average accuracy of 92.81% in the most favorable experiment, our proposed system showcases robust capability in spoken language identification tasks. Moreover, the system exhibits adeptness in learning language-specific patterns across various filter size representations of speech files, indicative of its versatility and efficacy in real-world applications.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134212

III. MATERIAL REQUIREMENTS

In the realm of spoken language identification, the initial phase commences with Requirement Elicitation, wherein stakeholders are engaged in interviews or surveys to elucidate their expectations and necessities for the system. This inclusive step fosters open discussions to grasp the system's purpose and its intended user base. Following this, Requirement Documentation ensues, encapsulating the gathered requirements in a structured manner through documents like use cases, user stories, or functional specifications. These documents serve as guiding principles for the development team and facilitate effective communication with stakeholders, laying a solid foundation for the subsequent phases.

Requirement Analysis ensues, delving deep into the gathered requirements to ensure clarity, completeness, and absence of ambiguities. High-level and detailed requirements are delineated, paving the way for a comprehensive understanding of the system's scope. Subsequently, Requirement Validation ensues, employing techniques such as requirement reviews, prototypes, and simulations to ensure alignment with stakeholders' needs and expectations.

Requirement Prioritization emerges as a pivotal step, guiding the determination of which features should be developed foremost, especially in Agile environments where iterative development is prevalent. Requirement Traceability is then established, ensuring that each requirement is meticulously tracked throughout the software development lifecycle, thereby facilitating change management and ensuring comprehensive coverage of all requirements.

The methodology systematically delineates the progression of developing, training, and evaluating a Convolutional Neural Network (CNN) tailored for language identification within homophonic short utterances. It initiates with meticulous dataset collection, ensuring a diverse array of homophonic samples across Telugu, Tamil, Hindi, Urdu, and Marathi languages, crucial for robust training and evaluation. Subsequent data preprocessing encompasses normalization, noise reduction, and segmentation to establish a standardized dataset. Additionally, multi-channel audio recordings are converted to mono-channel WAV format to streamline processing. Acoustic features extraction identifies key patterns and characteristics pivotal for CNN-based language identification. A custom CNN architecture is meticulously designed, optimizing layers and parameters to effectively capture intricate acoustic details inherent in homophonic utterances. Supervised learning empowers the model to discern language patterns amidst homophonic contexts. Validation and fine-tuning phases enhance accuracy and assess generalization capability. Performance evaluation metrics, including accuracy, precision, recall, and F1 score, gauge model effectiveness. Comparative analyses against traditional methods, coupled with optimization strategies like batch size adjustment and learning rate optimization, further refine model efficiency. Cross-validation techniques ensure robustness across dataset subsets, while real-time testing validates practical effectiveness. A comprehensive results analysis delineates strengths and limitations of the CNN-based approach in addressing homophonic challenges, affirming its efficacy in language identification tasks..

IV. METHODOLOGY

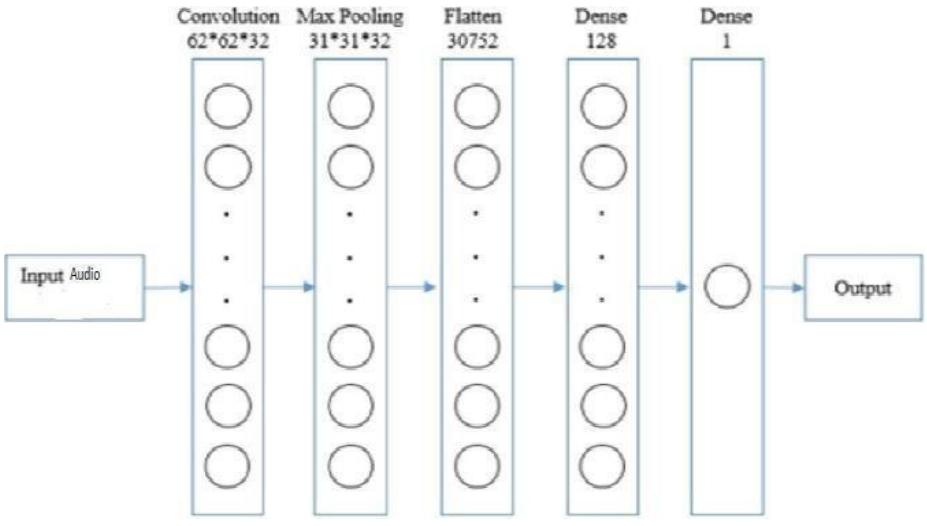


Figure 1. Language identification using 5-Layer Convolutional Neural Network

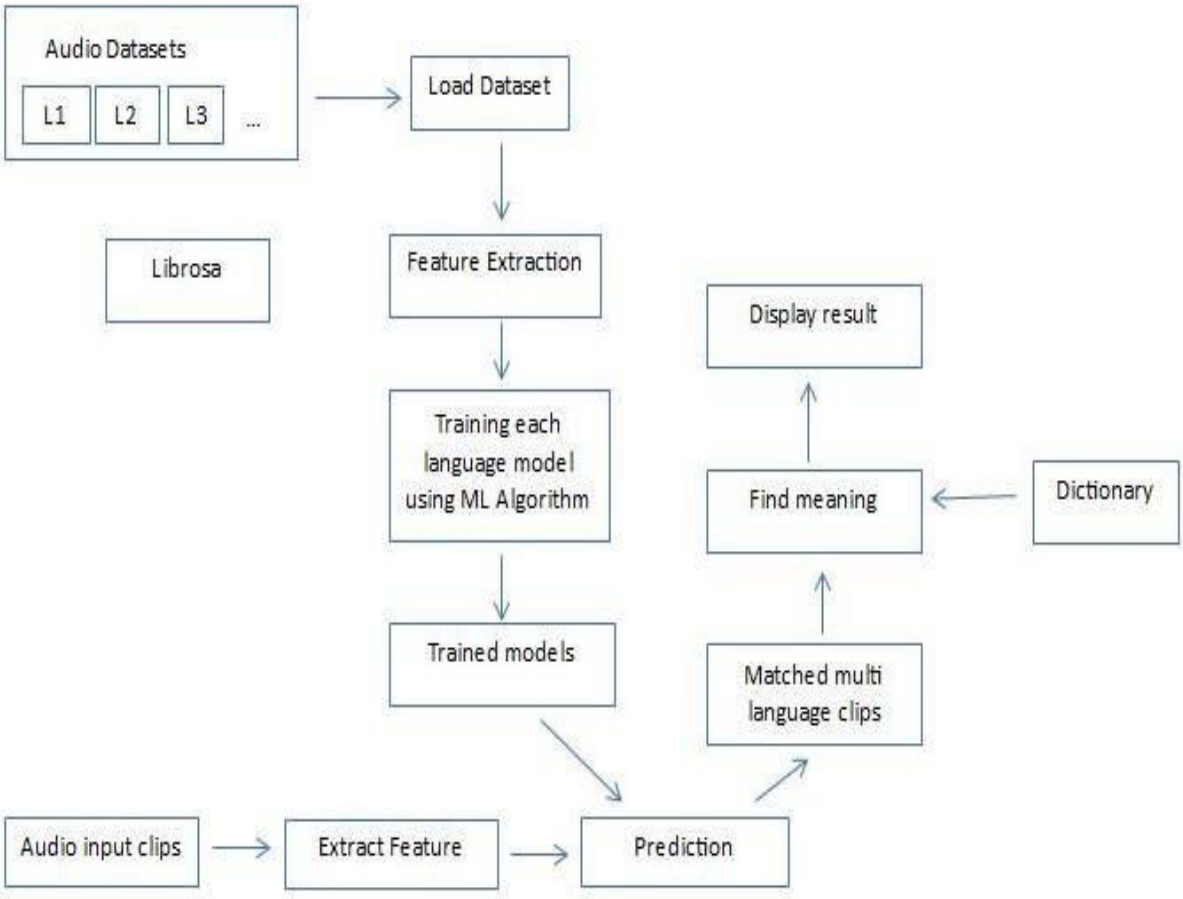
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134212

➢ **Proposed Design Architecture**



This block diagram depicts a process designed for audio data analysis, likely aimed at tasks such as language identification or classification. Initially, the audio datasets labeled as L1, L2, L3, etc., are loaded into the system for processing. Subsequently, Librosa, a Python package specialized in music and audio analysis, is employed for feature extraction from the audio data.

These extracted features serve as the basis for training individual language models, utilizing a Convolutional Neural Network (CNN) algorithm. Once trained, these models are then utilized for prediction or classification of audio data. Following the prediction step, there's an indication of an action to find meaning from the predicted results.

The resulting analysis is displayed, possibly in a user-friendly format for interpretation. Moreover, there appears to be a phase for matching audio clips across multiple languages, suggesting a multilingual aspect to the analysis. Ultimately, the outcomes of this analysis are stored in a CSV (Comma- Separated Values) file, enabling easy access and further processing of the results.

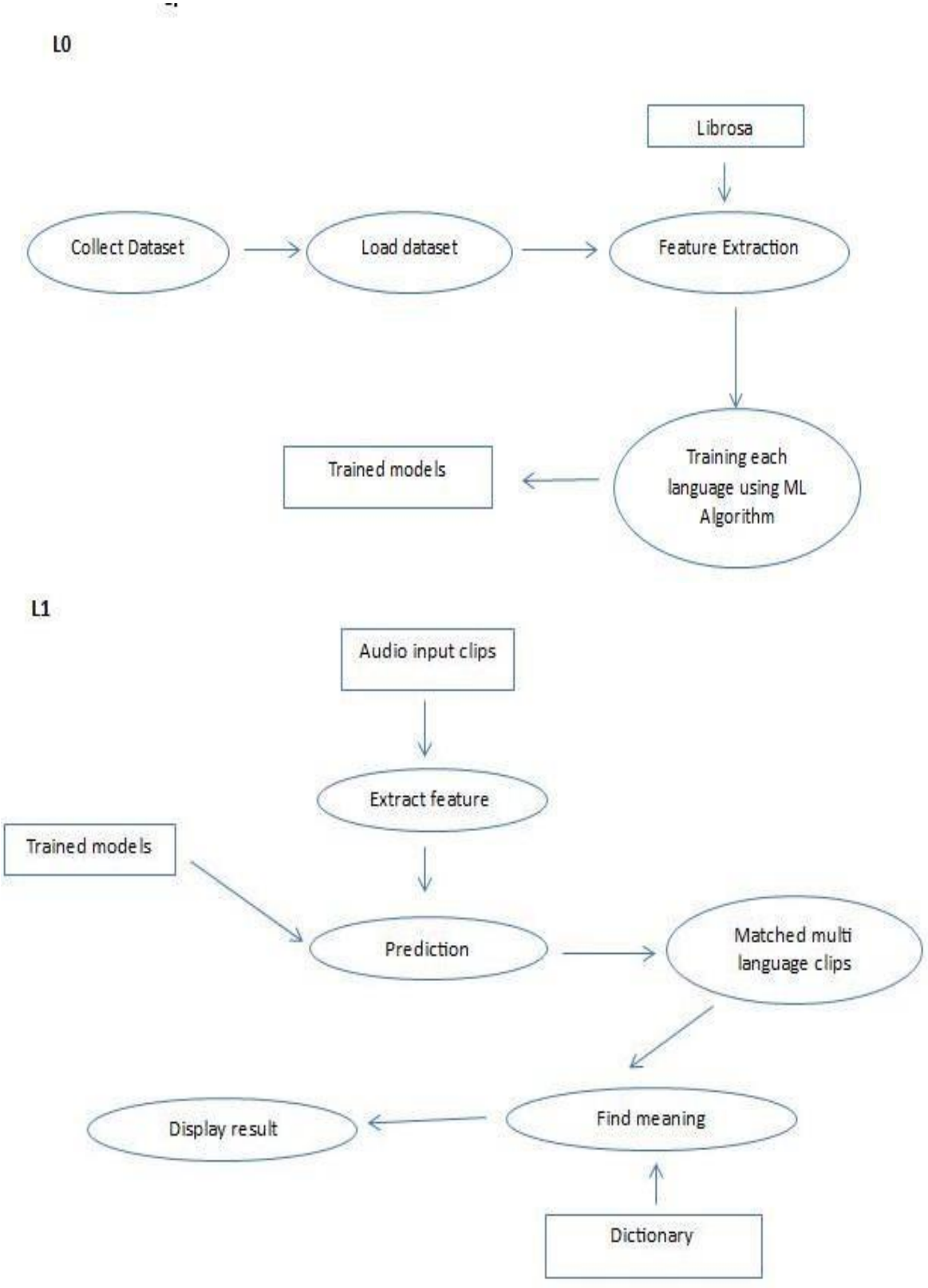
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134212

➢ **Data flow Diagram.**



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

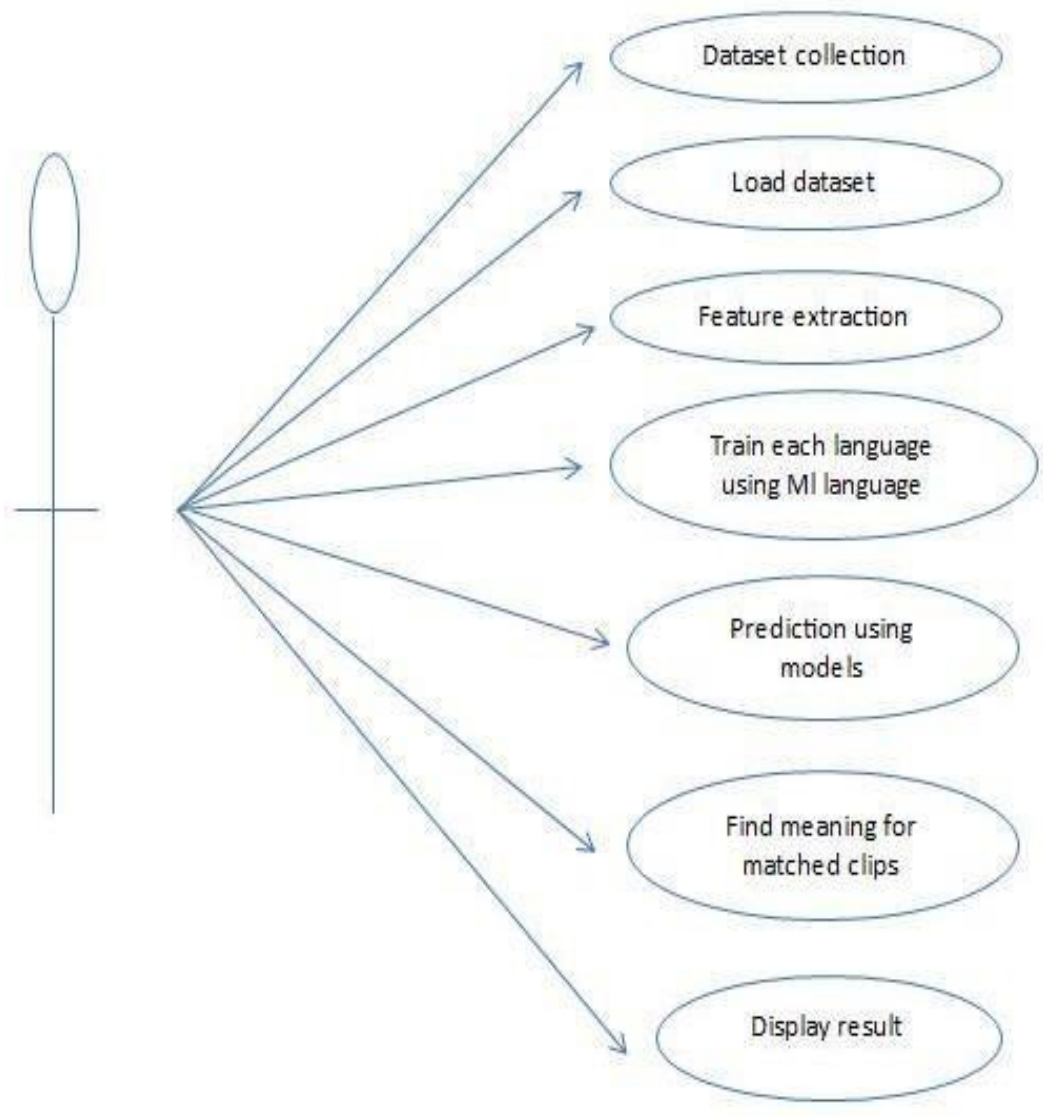
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134212

The dataflow diagram depict a comprehensive process likely involved in language identification or classification using CNN techniques applied to audio data. The level 1 outlines the initial steps, including collecting a dataset of audio samples, loading the dataset, and extracting features using the Librosa library. These features are then used to train models for each language using the Convolutional Neural Network (CNN) algorithm. Once trained, the models are obtained as the output. The level 1 builds upon this process, illustrating how trained models are utilized in a practical application. Audio input clips are fed into the system, where features are extracted, and predictions are made using the trained models. The results are then displayed, and further analysis, such as finding the meaning of the identified languages, can be performed, with outputs possibly saved in CSV files. Together, these flowcharts outline a systematic approach to language analysis, from initial data processing to practical application, demonstrating the integration machine learning in linguistic tasks.

- **Use Case Diagram.**



This diagram illustrates a process likely involved in language identification or classification, with a focus on user interaction and system response. The process begins with the user, who interacts with the system, possibly providing input data or initiating the analysis.The system then proceeds through several stages: dataset collection, loading the dataset, feature extraction, training each language using CNN language models, prediction using the trained models, finding meaning for matched clips, and displaying the results. Each stage involves specific actions taken by the system to process the input data and generate meaningful outcomes. Overall, this flowchart highlights the user's role in initiating the language analysis process and the system's subsequent steps to produce relevant results.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134212

ASSUPTION AND DEPENDENCIES

The proposed methodology for language identification within homophonic short utterances is built upon several key assumptions. Firstly, it assumes the existence of homophonic utterances in the multilingual speech dataset, where different languages produce similar sounds, posing a specific challenge that necessitates specialized techniques. Secondly, the success of the Convolutional Neural Network (CNN) model hinges on the assumption of reasonably clear and high-quality audio recordings, as noises or distortions might adversely affect its performance. Thirdly, the methodology assumes the availability of a representative dataset containing labeled homophonic utterances in multiple languages, crucial for training a supervised machine learning model like CNN. Several dependencies underpin the feasibility and efficacy of the proposed approach. Firstly, the success of the methodology depends on the availability of a diverse and comprehensive dataset containing homophonic utterances in multiple languages. Accurate and consistent labeling of this dataset is also essential, as the model heavily relies on correctly annotated data for training and testing. Moreover, sufficient computational resources, including processing power and memory, are crucial for efficient model training and evaluation. Compatibility with specific software libraries and frameworks, as well as reliable tools or algorithms for feature extraction, are also critical dependencies for the implementation of the proposed CNN- based approach.

EXPERIMENTATION

Experimentation within the project unfolds through a systematic process starting with the collection and preprocessing of a diverse dataset encompassing homophonic short utterances across multiple languages. The dataset is partitioned into training, validation, and test sets to facilitate model development and evaluation. Through iterative experimentation, various CNN architectures are designed, exploring different layer configurations, filter sizes, and activation functions to optimize performance. Hyperparameter tuning further refines the model by experimenting with parameters such as learning rate, batch size, and regularization strength. During training and validation phases, performance metrics such as accuracy, loss, precision, recall, and F1 score are monitored to assess model convergence and effectiveness. Subsequent to the experimentation process, evaluation is conducted on the test dataset to provide insight into the model's generalization capability and identify areas for improvement Comparative analysis against traditional methods and alternative architectures offers valuable benchmarks for assessing the CNN- based approach's performance. Furthermore, optimization strategies, including different algorithms and regularization techniques, are explored to enhance model efficiency. Cross-validation techniques validate model robustness, while real-time testing ensures practical effectiveness in real-world scenarios.

The experimentation process is driven by iterative refinement, with insights gained from each experiment informing subsequent iterations This iterative approach leads to continuous improvement, ultimately enhancing language identification accuracy and reliability in homophonic scenarios. Through systematic experimentation and refinement, the CNN-based language identification model evolves to achieve higher levels of performance and effectiveness in distinguishing languages within homophonic short utterances.

TESTING

Testing is an indispensable phase in the product development lifecycle, serving as a critical checkpoint to detect and rectify any remaining errors. This phase assumes paramount importance in ensuring the quality and reliability of the software, thus contributing significantly to the overall quality assurance process. Through a series of meticulously designed test cases, the program undergoes rigorous evaluation to ascertain its performance and adherence to expected behavior. Any errors identified during testing are duly corrected, and the corrective actions are meticulously recorded for future reference, thereby ensuring continuous improvement and refinement of the system before its implementation.

At its core, software testing is a technical investigation aimed at assessing the correctness, completeness, security, and overall quality of the developed software. It serves the stakeholders by providing quality-related information about the product within the context of its intended operation. However, it's essential to recognize that testing, while critical, cannot establish the absolute correctness of software; instead, it provides a comparative analysis against specifications, thus offering valuable insights into the product's state and behavior. It's imperative to distinguish software testing from Software QualityAssurance (SQA), which encompasses broader business process areas beyond testing.

Effective software testing requires a thorough investigative approach rather than merely following routine procedures. While there are various approaches to software testing, testing complex products necessitates dynamic analysis, putting the product through its paces to uncover potential issues. .quality attributes such as capability, reliability, efficiency, portability, maintainability, compatibility, and usability are integral considerations in the testing process. Ultimately, a good test is not just one that reveals errors but one that provides meaningful information to relevant stakeholders within the project community, thereby driving continuous improvement and ensuring the delivery of a high-quality software product.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134212

V. CONCLUSION

UIn conclusion, the project has successfully developed a Convolutional Neural Network (CNN)-based language identification system for homophonic short utterances. Through meticulous dataset collection, preprocessing, and model training using Keras, the system demonstrates promising accuracy in identifying languages even in challenging scenarios. Real-time testing validates its practical effectiveness across various applications. Moving forward, continuous refinement and exploration of advanced techniques will further enhance its applicability in multilingual speech processing.

ACKNOWLEDGMENT

We would like to thank the University that provided the necessary resources "**Dayananda Sagar University**" for the opportunities to make the study practical and accessible as desired.

We also thank professor "**Prof Sharath H A**" and friends for their collaboration, discussions, and helpful suggestions during a research conference and workshop where preliminary results of this study were presented. Their insights enrich our understanding and inspire us to renew our ways.

REFERENCES

[1]. Li, H., Ma, B., & Lee, K. A. (2013). Spoken Language Recognition: From Fundamentals to Practice. IEEE Transactions on Audio, Speech, and Language Processing, 21(5), 1088-1109.
[2]. Waibel, A., Geutner, P., Tomokiyo, L. M., Schultz, T., & Woszczyna, M. (2000). Multi linguality in Speech and Spoken Language Systems. IEEE Transactions on Speech and Audio Processing, 8(6), 684-695.
[3]. He, K., Xu, W., & Yan, Y. (2020). Multi-Level Cross-Lingual Transfer Learning With Language Shared and Specific Knowledge for Spoken Language Understanding. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 862-875.
[4]. Qamhan, M. A., Altaheri, H., Meftahi, A. H., Muhammad, G., & Alotaibi, Y. A. (2021). Digital Audio Forensics: Microphone and Environment Classification Using Deep Learning. IEEE Transactions on Information Forensics and Security, 16, 2991-3006.
[5]. Guha, S., Das, A., Singh, P. K., Ahmadian, A., Senu, N., & Sarkar, R. (2020). Hybrid Feature Selection Method Based on Harmony Search and Naked Mole-Rat Algorithms for Spoken Language Identification from Audio Signals. IEEE Transactions on Cybernetics, 50(6), 2613- 2625.
[6]. Shen, P., Lu, X., Li, S., & Kawai, H. (2020). Knowledge Distillation- Based Representation Learning for Short-Utterance Spoken Language Identification. IEEE Transactions on Audio, Speech, and Language Processing, 28, 860-869.
[7]. Padi, B., Mohan, A., & Ganapathy, S. (2020). Towards Relevance and Sequence Modeling in Language Recognition. IEEE Transactions on Audio, Speech, and Language Processing, 28, 999-1012.
[8]. Draghici, A., Abeßer, J., & Lukashevich, H. (2020). A Study on Spoken Language Identification using Deep Neural Networks. IEEE Transactions on Neural Networks and Learning Systems, 31(10), 4048-4062.

