Audio segmentation and Language Identification in Spanish and Highland Puebla Nahuatl

Anonymous ACL submission

Abstract

This paper investigates language identification in multilingual audio using Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) models. Our study highlights the superior performance of the CNN model with spectrogram data. However, challenges arise in distinguishing languages with significant overlap, as evidenced by confusion between 'azz' and 'esp' labels in short-duration spectrograms. Future work will explore training CNN models on longer frames to address these challenges. These findings contribute to advancements in language identification methodologies, particularly in code-switching contexts.

Keywords: Convolutional Neural Networks (CNN, ConvNets), Recurrent Neural Networks (RNN), Segmentation and Language Identification (SLID)

1 Introduction

In today's technologically advanced and automated world, the significance of speech recognition cannot be emphasized enough. From voice assistants like Alexa and Google assistant to the integration of voice-activated commands and audio transcription in platforms such as Microsoft Teams, along with the automatic captioning on YouTube, speech recognition plays a pivotal role. Given that nearly half of the global population is bilingual (Ansaldo et al., 2008; Byers-Heinlein et al., 2019), individuals often fluidly switch between languages in everyday conversations, making Language Segmentation as another crucial aspect of speech recognition, focusing on identifying the language boundaries within multi-language utterances.

This paper presents the implementation of Language Identification and Segmentation for Spanish and Highland Puebla Nahuatl audios. The motivation behind this research extends beyond the abstract uses of language segmentation; it seeks practical applications, particularly in the realm of managing bilingual radio transmissions. As part of our broader vision, we aim to contribute to the enhancement of radio archival search capabilities. The immediate use case revolves around enabling people to efficiently find programs by searching through transcriptions. The proposed process involves obtaining audio, performing language identification and segmentation, followed by automatic speech recognition (ASR) on the identified segments. This sequential approach lays the foundation for building a comprehensive database structured with key information such as date, time, language, program, segment begin, segment end, and corresponding text. This database, once established, can be easily searched and indexed, opening up new possibilities for efficient and accurate retrieval of information from bilingual radio transmissions.

2 Literature Review

The task of Language Identification has been of interest for quite some time. The early studies in this area involved exploring features such as Mel-frequency cepstral coefficients (MFCC), Shifted-DeltaCepstral (SDC), Perceptual linear prediction coefficient (PLP) and Linear Predictive Coding (LPC) (Bhattacharjee and Sarmah, 2013; Sarmah and Bhattacharjee, 2014; Moselhy and Abdelnaiem, 2013). (Kumar et al., 2010) trained Vector Quantization (VQ) with Dynamic Time Warping (DTW) and Gaussian Mixture Model (GMM) classifiers on MFCC , PLP, along with two hybrid features, Bark Frequency cepstral coefficient (BFCC) and Revised Perceptual Linear Prediction coefficient (RPLP) for language identification. Later, (Ali et al., 2015) investigated the performance of generative models such as n-gram language model, Naive Bayes as well as discriminative models like linear SVM and Maximum Entropy using Lexical, Phonetic (n-gram phone sequence) and i-vector features on a

related task of dialect detection.

The later works in this field witnessed a shift from the vector space approach to the deep learning approach. After observing the effective implementation of DNNs in acoustic modeling for speech recognition, (Gonzalez-Dominguez et al., 2015) applied DNNs to the task of language detection utilizing short-term acoustic characteristics of an utterance. In their work, they highlighted the ability of DNNs to generate language identification posterior for each new frame of the test utterance, thus making them suitable for real-time LID task. Motivated by the ability of RNNs to model sequential data, (Gonzalez-Dominguez et al., 2014; Zazo et al., 2016) implemented RNNs for LID and showed that they can effectively exploit temporal dependencies in acoustic data and learn relevant features for language discrimination purposes. (Zazo et al., 2016) pointed out that most DNN and RNN based models overlook the phonetic information of the utterance. In their work, they implemented RNN-LSTM model that used frame level phonetic features produced by a phone-discriminative DNN rather than raw acoustic features. In order to design effective representation for LID that is robust to speaker, channel variability, and background noise, (Jin et al., 2017) extracted first and second-order statistics from a phoneme-related DNN and stacked it by convolutional layers. Similar works by (Wang et al., 2019; Chakraborty et al., 2021; Singh et al., 2021) explored different CNN architectures along with techniques like data augmentation and audio transformation that improved the performance of CNN models for this task. Another interesting work by (Draghici et al., 2020) involved hierarchical classification using CNN where, first, Language family is identified followed by the identification of language.

More recent approaches focuses on exploiting a CNN-RNN hybrid architecture. One such implementation by (Bartz et al., 2017) operates on spectrogram images of the audio snippets where CNN captures spatial information and RNN captures information through sequence of time steps. For making the model robust to noises, he added three types of noises - (i) randomly generated white noise, (ii) periodic cracking noise (emulating bad voice chat connection), and (iii) background music from different genres. In a similar work, (Sarthak et al., 2019) highlighted that using phonemes based prosodic and acoustic features for languages that have overlapping phonemes makes the classification task challenging. They proposed a CRNN based attention model which uses log-Mel spectrogram image for implementing SLID models for such languages.

Language	Number of audios	Total hours
Spanish	8057	
Nahuatl	8531	
Both languages	356	
Neither	3737	
Total	20861	

Table 1: This table summarizes the number of audios in each category for the dataset used in this study

3 Dataset

The dataset utilized in this study originates from the bilingual Nahuatl-Spanish news program, Ejekat Tanauatilme, produced by the Radio Tsinaka station [1] in San Miguel Tzinacapan, Puebla, Mexico. This program, released multiple times per month, follows a distinctive format where each episode incorporates a segment covering a story predominantly in Nahuatl, succeeded by the same segment presented primarily in Spanish. Additionally, there are instances of both languages being used within a single segment, exemplifying intrasentential code-switching. To compile the dataset, we collected a set of 24 hour-long episodes from the radio station.

To facilitate further analysis, the audio data was automatically segmented into utterances using ELAN's built-in "Fine audio segmenter." This segmentation process was instrumental in breaking down the episodes into more manageable units for subsequent annotation. The resulting utterances varied in duration, spanning from 1 to 20 seconds, generating a substantial collection of 20,681 audios. To establish the language classification of each utterance, a fluent Nahuatl speaker meticulously listened to the clips and annotated them using ELAN. The annotations categorized each utterance as either (a) Spanish, (b) Nahuatl, (c) both languages, or (d) neither (e.g., non-verbal segments such as music playing). This comprehensive annotation process serves as a crucial foundation for the subsequent language identification and segmentation tasks undertaken in this research. Table 1 summarizes the statistics of the audios.

[1] https://radiotsinaka.org/

4 Method

The initial step involved the transformation of raw audio data and corresponding annotations into a requisite format for model training. Subsequently, an architecture based on Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN)[2] was devised, and was trained within a supervised learning framework with an objective to predict the language from a transformed input data.

4.1 Data Transformation

Two distinct transformations were applied to the raw audio data to prepare it for model training. First, spectrogram images were generated for each audio sample utilizing the specgram method from the SciPy library. These spectrogram images represent the frequency spectrum of the audio signal, with time delineated on the x-axis (seconds) and frequency on the y-axis (Hz). Second, Mel-frequency cepstral coefficients (MFCC) were extracted from the audio samples using the mfcc method from the Librosa library. This process yielded 13 MFCC coefficients for each audio sample.

To facilitate the classification task, the language labels ('na', 'azz', 'esp', 'ambos') were mapped into categorical labels using a predefined mapping {'ambos': 0, 'azz': 1, 'esp': 2, 'na': 3}. Subsequently, these labels were encoded into categorical format to align with the model's requirements.

The transformed data was partitioned into training, validation, and test sets in an 80:10:10 ratio to ensure adequate training, validation, and evaluation of the models.

4.2 Data Modelling

In our research, we aimed to train a robust language identification model and then further use it for the language segmentation task.

4.2.1 Language Segmentation Task

For Language Identification Task, we experimented with CNN and RNN based model architecture. Both CNN and RNN models were trained using their respective input data types, aiming to learn discriminative features and accurately classify the language categories based on the transformed audio data. Figure 1 and Figure 2 depict the process flow of generating predictions using a CNN and an RNN model, respectively.

[2] https://github.com/nsingh475/Segmentation_and_Language_Identification

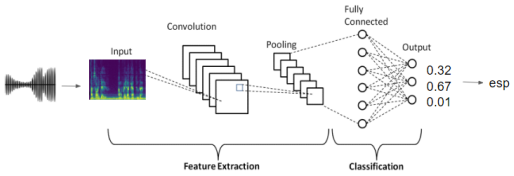

Figure 1: A flow chart demonstrating language prediction using CNN model

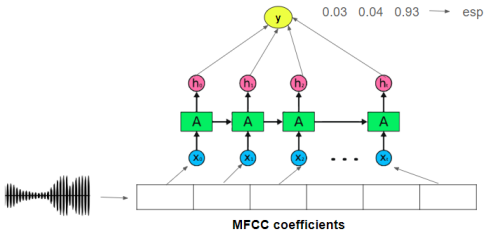

Figure 2: A flow chart demonstrating language prediction using RNN model (with MFCC coefficients)

CNN Model: The CNN architecture was designed to operate on the spectrogram images generated from the audio data. The model consisted of multiple convolutional layers with rectified linear unit (ReLU) activation functions, followed by max-pooling layers for spatial downsampling. Additionally, dropout regularization was applied to mitigate overfitting. The flattened feature maps were then passed through fully connected dense layers with ReLU activation functions. The output layer utilized softmax activation to classify the input into one of the four language categories.

RNN Models: Two variants of RNN models were developed to leverage different representations of the audio data. The first model utilized the MFCC coefficients extracted from the audio samples. It comprised bidirectional Long Short-Term Memory (LSTM) layers with varying units. Batch normalization was incorporated after each LSTM layer to stabilize and accelerate training, while dropout regularization was employed to prevent overfitting.

The second RNN model utilized sequences of spectrogram images as input data. This model shared a similar architecture with the MFCC-based model, featuring bidirectional LSTM layers, batch normalization, and dropout regularization. However, the input shape was configured to accommodate the sequential nature of spectrogram images.

4.2.2 Language Segmentation Task

In the language segmentation task, we aimed to identify the span of language in a code switching audio setup, specifically targeting audios annotated as 'ambos' in the original dataset. To achieve this, the input audio was partitioned into smaller temporal frames, typically lasting 1 or 2 seconds each. Subsequently, a spectrogram image was generated for each frame. These spectrograms served as input data for the CNN model, which predicted the language label for each frame. The aggregated labels from all frames were then utilized to construct the comprehensive sequence of language labels corresponding to the input audio. Figure 3 presents the process of language segmentation using the trained CNN model.

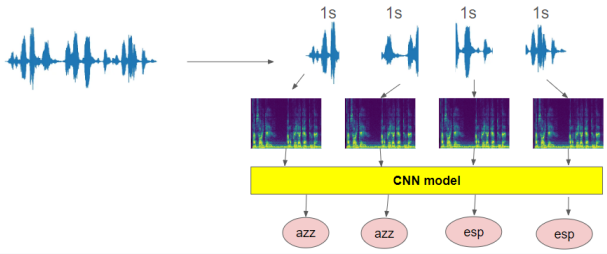

Figure 3: A flow chart demonstrating language segmentation using CNN model (with 1s frame)

5 Results

Table 2 presents a summary of the experimental results of the trained CNN and RNN models fro the language Identification task. The results indicate that the CNN model achieved the highest accuracy of 99.30% when trained on spectrogram images. However, when the RNN model was trained on spectrogram sequences, the accuracy slightly decreased to 92.60%. Notably, using MFCC coefficients as input data for the RNN model resulted in a lower accuracy of 82.19%.

Table 3 presents the results of the Language Segmentation task on monolingual audio data. The objective of this task was to evaluate the performance of a CNN model for language segmentation task prior to its application in a code-switching context. In the code-switching scenario, denoted as 'ambos', the audios contain utterances from both the Spanish ('esp') and Highland Puebla Nahuatl ('azz') languages. Table shows that 1s frame duration, the accuracy of 'azz' is high for both single-frame and multi-frame setup but very low for 'esp' and 'na'. Upon manual inspection, it was observed that the

Model	Data	Accuracy
CNN	Spectrogram	99.30%
RNN	Spectrogram Sequence	92.60%
RNN	MFCC Coefficients	82.19%

Table 2: This table reports the accuracy of CNN and RNN models on the Language Identification task.

1s duration signal proved insufficient for accurate language identification, leading to a predominant classification of 'azz' for the majority of the audios. To address this issue, we employed two strategies: (i) replicating the 1s signal four times prior to spectrogram creation, and (ii) employing a 2s frame for spectrogram generation. Both of these strategies resulted in improved accuracy.

Duration	frame	azz	esp	na
1s	single	74.27	4.19	1.72
	multiple	87.5	19.65	47.89
1s_4x	single	50.43	45.35	42.96
	multiple	54.31	58.67	2.82
2s	single	51.52	49.12	43.04
	multiple	21.74	52.17	88.68

Table 3: This table presents a summary of the Language Segmentation task results, showcasing the accuracy percentages for both single-frame and multi-frame audios, analyzed separately. The experiments encompassed frames of 1s and 2s duration. Specifically, '1s_4x' indicates a 1s frame repeated four times before generating the spectrogram.

6 Conclusion

The findings of language identification task suggests that the CNN model performed exceptionally well with spectrogram data, while the RNN model exhibited slightly lower accuracy, especially when trained on MFCC coefficients. Moreover, the language identification task revealed challenges in the model's ability to discern distinctive representations for languages exhibiting substantial overlap. Notably, the model exhibited notable confusion between 'azz' and 'esp' labels when presented with spectrograms of 1s or 2s frames.

7 Future Work

Moving forward, our aim is to train a CNN model using a 5s frame for the language identification task. Subsequently, we intend to utilize this trained model to extract language segments within a code-switching context.

References

Ahmed Ali, Najim Dehak, Patrick Cardinal, Sameer Khurana, Sree Harsha Yella, James Glass, Peter Bell, and Steve Renals. 2015. Automatic dialect detection in arabic broadcast speech. *arXiv preprint arXiv:1509.06928*.

Ana Inés Ansaldo, Karine Marcotte, Lilian Scherer, and Gaelle Raboyeau. 2008. Language therapy and bilingual aphasia: Clinical implications of psycholinguistic and neuroimaging research. *Journal of Neurolinguistics*, 21(6):539–557.

Christian Bartz, Tom Herold, Haojin Yang, and Christoph Meinel. 2017. Language identification using deep convolutional recurrent neural networks. In *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14–18, 2017, Proceedings, Part VI 24*, pages 880–889. Springer.

Utpal Bhattacharjee and Kshirod Sarmah. 2013. Language identification system using mfcc and prosodic features. In *2013 International Conference on Intelligent Systems and Signal Processing (ISSP)*, pages 194–197.

Krista Byers-Heinlein, Alena G Esposito, Adam Winsler, Viorica Marian, Dina C Castro, and Gigi Luk. 2019. The case for measuring and reporting bilingualism in developmental research. *Collabra: Psychology*, 5(1):37.

Neelotpal Chakraborty, Soumyadeep Kundu, Sayantan Paul, Ayatullah Faruk Mollah, Subhadip Basu, and Ram Sarkar. 2021. Language identification from multi-lingual scene text images: a cnn based classifier ensemble approach. *Journal of Ambient Intelligence and Humanized Computing*, 12:7997–8008.

Alexandra Draghici, Jakob Abeßer, and Hanna Lukashevich. 2020. A study on spoken language identification using deep neural networks. In *Proceedings of the 15th International Audio Mostly Conference*, pages 253–256.

Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, Pedro J Moreno, and Joaquin Gonzalez-Rodriguez. 2015. Frame-by-frame language identification in short utterances using deep neural networks. *Neural Networks*, 64:49–58.

Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, and Hasim Sak. 2014. Automatic language identification using long short-term memory recurrent neural networks.

Ma Jin, Yan Song, and Ian Vince McLoughlin. 2017. End-to-end dnn-cnn classification for language identification.

Pawan Kumar, Astik Biswas, Achyuta Nand Mishra, and Mahesh Chandra. 2010. Spoken language identification using hybrid feature extraction methods. *arXiv preprint arXiv:1003.5623*.

Abdallaa Mohammed Moselhy and Abdelaziz Alsayed Abdelnaiem. 2013. Lpc and mfcc performance evaluation with artificial neural network for spoken language identification. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 6(3):55.

Kshirod Sarmah and Utpal Bhattacharjee. 2014. Gmm based language identification using mfcc and sdc features. *International Journal of Computer Applications*, 85(5).

Sarthak, Shikhar Shukla, and Govind Mittal. 2019. Spoken language identification using convnets. In *Ambient Intelligence: 15th European Conference, AmI 2019, Rome, Italy, November 13–15, 2019, Proceedings 15*, pages 252–265. Springer.

Gundeep Singh, Sahil Sharma, Vijay Kumar, Manjit Kaur, Mohammed Baz, and Mehedi Masud. 2021. Spoken language identification using deep learning. *Computational Intelligence and Neuroscience*, 2021.

Yutian Wang, Huan Zhou, Zheng Wang, Jingling Wang, and Hui Wang. 2019. Cnn-based end-to-end language identification. In *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, pages 2475–2479. IEEE.

Ruben Zazo, Alicia Lozano-Diez, Javier Gonzalez-Dominguez, Doroteo T. Toledano, and Joaquin Gonzalez-Rodriguez. 2016. Language identification in short utterances using long short-term memory (lstm) recurrent neural networks. *PloS one*, 11(1):e0146917.