Enhancing Audio Data Processing: Insights from the Development and Evaluation of a Transcriber tool

Carlo A. Castro
University of the Immaculate Conception
Davao, Davao del Sur, Philippines
ccastro_200000000215@uic.edu.ph

Muslimin B. Ontong
University of the Immaculate Conception
Davao, Davao del Sur, Philippines
montong_200000000677@uic.edu.ph

Aurora Cristina Manseras
University of the Immaculate Conception
Davao, Davao del Sur, Philippines
amanseras@uic.edu.ph

Kristine Mae M. Adlaon
University of the Immaculate Conception
Davao, Davao del Sur, Philippines
kadlaon@uic.edu.ph

ABSTRACT

Language classification models play a crucial role in various natural language processing applications, including machine translation. While significant research has been conducted on text-based language classification, relatively less attention has been given to audio data. This paper aims to bridge this gap by exploring the development of a tool specifically designed for classifying audio inputs, with a particular focus on indigenous languages. The Minna Transcriber Tool is a solution tailored to preserve audio files and extract features by generating metadata for each audio segment. Additionally, the paper delves into the creation of a language classification algorithm capable of accurately identifying indigenous languages from audio recordings. Through a combination of classical machine learning techniques and deep learning algorithms.

KEYWORDS

Datasets, neural networks, natural language processing, translation

1 INTRODUCTION

Languages represent a cornerstone of human diversity, serving as conduits through which we perceive, interact with, and interpret the world in distinct ways. They encapsulate our cultures, collective memories, and values, forming an integral part of our identities [1]. In today's interconnected world, improving communication through technology is crucial [14]. By using technology to enhance communication, individuals, businesses, and organizations can overcome barriers and build stronger relationships [11]. One of the significant endeavors in language preservation in the Philippines is the work of Department of Science and Technology, which primarily concentrates on developing tools and technologies for Mindanao languages [4]. A work titled 'SultiWag' have collected more than three thousand (3,000) words from the combined Manobo, Kagan, and Davaoeño-Cebuano languages. Additionally, seventy-five (75) raw recordings were produced in three (3) languages during recording sessions. The data went through four (4) phases in preprocessing: Cutting, Extraction, Conversion, and Trimming, then files were saved in mp3 format. The Sultiwag's researchers converted the audio into spectrograms and exported them as JPEG files. The process resulted to 70% language classification accuracy. Primarily, the objective of this study is to explore an alternative method of data pre-processing and data feeding using spectrogram data directly, instead of saving it as JPEG. This approach is chosen to mitigate potential data loss during image export and to enhance the accuracy of audio classification, particularly for indigenous languages.

Indigenous languages serve a deeper purpose beyond communication—they act as a bridge to a community's heritage, nurturing a profound sense of belonging. These languages also act as vessels, carrying the ethical principles passed down through generations, shaping the values of those who speak them. Indigenous languages are disappearing worldwide, but there are numerous efforts and notable achievements in safeguarding these Indigenous languages and culture [12].

Language has the potential to thrive and endure if there is collective effort. With adequate support and resources at various levels – transnational, national, local, community, and individual, these dying languages can be revitalized or preserved [7]. Another study highlighted the ongoing efforts to preserve the diverse Indigenous languages of the Philippines. The Department of Education initiated a program called the Mother Tongue-based Multilingual Education program, aiming to revive and preserve Mother Tongue languages [8].

While Artificial Intelligence offers exciting possibilities for language revitalization, as seen in projects of [9], the key lies in collaboration. Indigenous communities must be active partners in developing AI tools for their languages. This ensures the technology respects the cultural context and avoids past exploitation by large corporations. Furthermore, such collaboration empowers communities to preserve their languages and traditions in the digital age. Imagine an AR experience where children learn Kwak'wala while navigating a virtual potlatch ceremony. AI, used thoughtfully and with respect, can bridge the digital divide and ensure these cultural treasures are not lost.

Another pivotal technology for language preservation and classification is audio data processing. It aids in pattern recognition by leveraging the distinctive acoustic attributes of various languages, thereby enhancing the efficacy of classification algorithms [3]. Moreover, audio data processing finds application in the creation of speech transcription systems for indigenous languages [13]. Techniques such as spectrograms and signal processing have proven instrumental in extracting crucial features from audio signals [2, 10].

Spectograms are graphs of audio signals that present carrier frequency and intensity change over time [5]. These kinds of spectrogram representations help researchers analyze and compare language-specific characteristics, i.e., phonetics and prosody features [10]. It is a 3D representation where time is on the X-axis, frequency on the Y-axis, and frequency amplitude on the Z-axis. This visualization helps identify the significant features and common patterns within language audio signals which helps to have strong models for classification [10]. Working with audio data may be challenging, as the voice quality, background noise, and speaker variations significantly impact the results. Moreover, having no access to datasets that are collected in different languages and different accents could make this task more complicated. To overcome this, data augmentation techniques and robust feature extraction methods can be employed. One of such techniques includes a number of data augmentation methods such as time masking, pitch shifting and noise injection which used for generation of a larger scale of the data from which the models for classification of spoken languages should be learned [6].

2 METHODOLOGY

This study aimed to improve the efficiency and accuracy of language classification for indigenous languages, particularly the Manobo and Kagan datasets. The revisiting researchers developed audio dataset pre-processing to enhance the quality of the recordings, which involved removing background noise and ensuring the clarity and accuracy of spoken words.

2.1 Data Collection

The researchers used two particular subsets, 314 Manobo words, and 405 Kagan words, and collected a total of 2959 audio datasets. The use of such groups was undertaken since it made it possible to investigate the specific problems and understand better the problems to solve with language classification algorithms.

2.2 Data Analysis

A close check on the researcher's data preprocessing methods was done to guarantee the accuracy and reliability of the subsequent analyses. The researchers exhaustively identified numerous speakers, ranging from different accents and cohorts. They processed words spoken by speakers in the audio files through a free and open-source digital audio editor and recording application software to filter out background noise from the specific audio files. Subsequently, they marked and labeled the segmented audio using a consistent format:

> **(English Word)_(Indigenous Translation)(First character of language)**
>
> **For example: abandon_paguyow(M)**

This approach enabled the researchers to organize the data effectively, streamlining further analysis. To determine which features must be extracted and pre-processed before training the model, further examination of the data was conducted. Upon examination, the researchers discovered that certain attributes, such as sample rate and duration, required careful pre-processing and extraction from the audio files. Sample rate refers to the number of audio samples captured per second, determining the frequencies represented in digital audio. Meanwhile, duration is employed to standardize audio clips to a consistent length. As various sounds have distinct sample rates, re-sampling them to a common rate can aid in audio classification.

2.3 Data Pre-processing

The researchers opted to analyze a focused subset of the dataset, specifically Manobo (314 words) and Kagan (405 words), to enable a more in-depth analysis of these languages for classification. This focus, however, resulted in an imbalanced dataset. To address this, class weights were assigned, prioritizing the underrepresented language during model training.

In the preprocessing stage, the audio data underwent standardization to a uniform sample rate of 16,000 Hz. Additionally, researchers investigated the optimal fixed duration for the audio files by analyzing the mean and standard deviation of their lengths. Due to the limited number of audio samples, data augmentation techniques like time shift, time mask, and frequency mask were implemented to enrich the dataset's diversity. Finally, the preprocessed audio data was transformed into spectrograms, a visual representation capturing sound wave information across frequency, time, and intensity. Spectrograms, well-suited for training Convolutional Neural Networks (CNNs) in audio classification tasks, were then converted into a format compatible with the CNN model for further analysis.

2.4 Train Model

In the development of the language classification models, the researchers employed a specific process to train the models:

- **Loss Function:** The Sparse Categorical Crossentropy loss was utilized, suitable for multi-class classification tasks where the labels are integers.
- **Optimizer and Scheduler:** An Adam optimizer was applied with an exponential decay learning rate scheduler. The initial learning rate was set at 0.001, and it reduces by a factor of 0.9 every 1000 steps.
- **Model Checkpointing:** The model's state achieving the highest validation accuracy during the training was preserved. This allows for the use and further fine-tuning of the best-performing model snapshot.
- **Early Stopping:** An early-stopping mechanism was implemented to curtail overfitting. If the validation loss failed to improve for 10 consecutive epochs, the training process was halted, and the weights from the epoch with the lowest validation loss were reinstated.
- **Training:** The models were trained for several epochs, with class weights applied to the loss function to manage class imbalance. The models' progress was monitored using the accuracy metric.

To prevent overfitting, the researchers employed a technique called 5-fold cross-validation. This method splits the data into 5 parts, trains the model on 4 parts, and tests it on the remaining one. This process is repeated 5 times, ensuring the model's generalizability to unseen data.

2.5 Development

Part of the output of this research is the development of a desktop application called 'Minna Transcriber.' This tool was instrumental in facilitating the analysis of the existing dataset in a more efficient manner. It simplifies the file encoding process and reduces the likelihood of errors during file renaming. Moreover, the application was designed to export datasets in convenient formats such as CSV, JSON, and Excel.

During experiments, Google Colab was used to experiment with methods to enhance the accuracy of the machince learning model. Other development tools that were used include Flutter, for the creation of a high-fidelity prototype for the project, and Appwrite, for the back-end infrastructure.

3 PRELIMINARY RESULTS

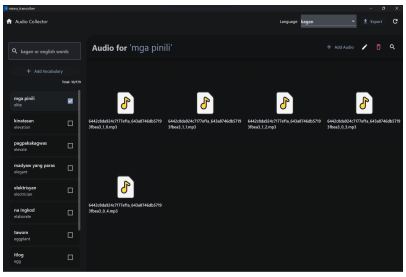

Figure 1: Minna Transcriber Tool

The development of "Minna Transcriber" resulted to streamlining the process of encoding and formatting audio files. This tool effectively prevents issues related to renaming file names while preserving features and target labels, resulting in a more efficient and user-friendly experience. Furthermore, this tool significantly contributed to our data analysis task before pre-processing the dataset.

3.1 Train Model

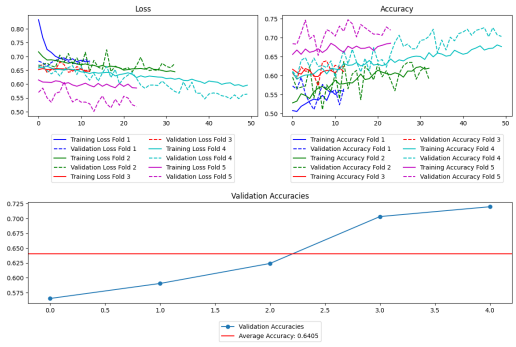

Figure 2: UBLCM001A Results

Looking at Figure 2, the researchers observed a pattern where the training and validation loss decrease while the training and validation accuracy increase as the number of epochs increases. The model was trained for a maximum of 100 epochs but didn't always reach that limit. The researchers implemented an early stop technique, meaning the training stops if the accuracy decreases continuously for five consecutive epochs. When this happens, the model is considered to have achieved good performance, and its weights and validation results are saved.

The model stops training for this specific k-fold training process when the accuracy declines for five consecutive epochs. Then, these saved weights and validation results are used for the next k-fold iteration. This process continues until all five folds are completed.

Figure 2 shows that the model typically reaches convergence around 50 epochs. The average validation accuracy at this point is 64%. If the researchers apply this trained model to new, unseen data (test dataset), it achieves an accuracy of 71%.

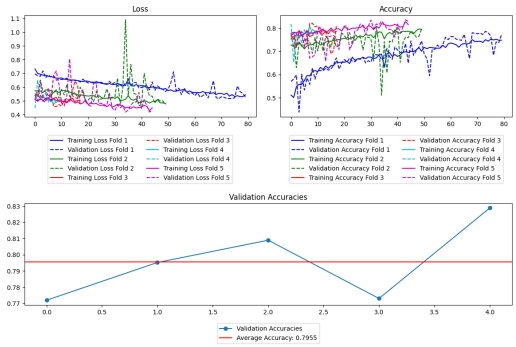

Figure 3: UBLCM002A Results

The results for the UBLCM002A model, as depicted in Figure 3, exhibit a similar trend to the UBLCM001A model. In all five k-folds, the model reached convergence after 80 epochs. On average, the validation accuracy at this point was 80%. When this trained model was evaluated on new, unseen data or test datasets, it achieved an accuracy of 82%.

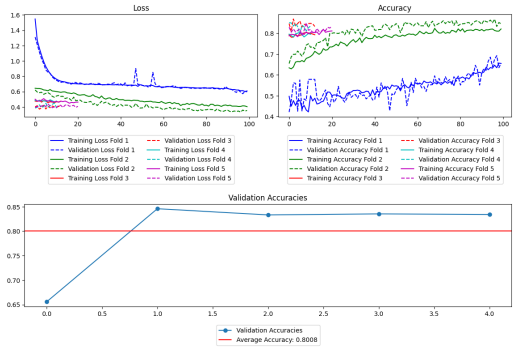

Figure 4: UBLCM005A Results

Referring to Figure 4, the trend observed in the model's performance is consistent with the previous models. However, a key distinction is that for nearly all five k-folds, this model could complete the entire 100 epochs. This indicates the potential for further