BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies

https://journals.bilpubgroup.com/index.php/fls

RESEARCH ARTICLE

A Multimodal Approach to Language Identification in Sotho-Tswana Musical Videos

Osondu Everestus Oguike [*] ***, Mpho Primus***

Institute for Intelligent Systems, University of Johannesburg, Johannesburg, South Africa

ABSTRACT

Language plays a crucial role in Sotho-Tswana musical videos, as it helps determine the sentiment and genre. The Sotho-Tswana languages, spoken in parts of Southern Africa, are used to compose many indigenous songs and music. However, speakers of one of the Sotho-Tswana languages may not understand other Sotho-Tswana languages. Given the widespread availability of these musical videos on social media platforms, there is a need for appropriate recommendations for users based on the language used in the videos. While traditional language identification in music has focused on audio, music information for identifying the singing language can also be embedded in other modalities, such as visual and text. This study employs a multimodal approach to identify the singing language in Sotho-Tswana musical videos. The multimodal approach focuses on three modalities, visual, audio, and textual/lyrics. A multimodal dataset of Sotho-Tswana musical videos is used to train deep learning and language models, for each of the modalities. After the independent training, for each of the modalities, a decision-level (late) fusion method is used to combine the results of the training from the three modalities. The results demonstrate that a multimodal approach outperforms single-modality methods, such as those relying solely on lyrics or textual information.

Keywords: Singing Language Identification; Multimodal; Audio Modality; Visual Modality; Textual Modality; Deep Learning

*CORRESPONDING AUTHOR:

Osondu Everestus Oguike, Institute for Intelligent Systems, University of Johannesburg, Johannesburg, South Africa; Email: osonduo@uj.ac.za or osondu.oguike@unn.edu.ng

ARTICLE INFO

Received: 2 October 2024 | Revised: 5 November 2024 | Accepted: 8 November 2024 | Published Online: 9 January 2025
DOI: https://doi.org/10.30564/fls.v7i1.7623

CITATION

Oguike, O.E., Priimus, M., 2025. A Multimodal Approach to Language Identification in Sotho-Tswana Musical Videos. Forum for Linguistic Studies. 7(1): 741–752. DOI: https://doi.org/10.30564/fls.v7i1.7623

1. Introduction

Automatic language identification is a task that determines the language used in spoken utterances and has been well-researched [1]. However, applying language identification to music information retrieval, specifically identifying the language used in music, remains underexplored [2]. The significance of the natural language used in musical videos in music information retrieval cannot be overstated, as it aids in understanding music sentiment [3, 4] and music genre [5].

Traditional approaches to identifying the language used in music have primarily focused on the audio modality [2]. However, other modalities, such as lyrics/text and visual elements, can provide supplementary information for language identification. For instance, Choi and ByteDance [6] combined audio and lyrics to enhance language identification.

Most research in this area has concentrated on Western music, with some studies identifying languages in non-Western music, such as Bangla and Hindi [7] and Indian languages [8]. Furthermore, language identification in music, for languages like Mandarin, English, and Japanese has also been explored [2].

Different techniques have been used to identify the singing language in music, among them includes: phoneme recognition, which is a deep phonotactic approach that employs Connectionist Temporal Classification for phoneme and language recognition [9]. Another technique is the ensemble learning technique, which uses Line Spectral Frequency-Approximation Gradation (LSF-AG) to achieve a high accuracy rate of 98.6% in identifying languages, like English, Bangla, and Hindi, in music [7].

Despite the extensive research on language identification in various musical contexts, there is a notable absence of a multimodal approach to language identification in Sotho-Tswana musical videos. This study aims to fill this gap by employing a multimodal methodology to identify the singing language in Sotho-Tswana musical content, this remains the novelty of this study.

1.1. Statement of the Problem

This study tackles several pressing issues related to Sotho-Tswana musical videos, focusing on language identification and its broader implications. The challenges addressed include the absence of language-based music recommenders, the under-resourced status of Sotho-Tswana languages, the limitations of unimodal language identification approaches, and the threat of language extinction.

- **Lack of Language-Based Sotho-Tswana Music Recommenders:**
  The proliferation of social media platforms for disseminating multimedia content, including Sotho-Tswana musical videos, necessitates the development of language-based music recommenders. Such recommenders would enable social media users to listen to and watch Sotho-Tswana musical videos that they understand, enhancing user experience and accessibility. Currently, no such tool exists for Sotho-Tswana musical videos.
- **Under-Resourced Languages:**
  Sotho-Tswana languages are considered under-resourced due to the limited availability of natural language processing resources, like datasets, compared to Western languages like English [10]. Developing a language identification tool for Sotho-Tswana musical videos represents a significant step towards increasing the natural language resources available for these languages, thereby shifting them from under-resourced to high-resourced status.
- **Non-Inclusiveness of Different Modalities:**
  Traditional language identification approaches for musical videos typically use an unimodal approach, relying solely on either audio [2]. However, effective language identification requires leveraging information available across multiple modalities, including audio, lyrics/text, and visuals. The exclusion of any modality can result in suboptimal language identification for Sotho-Tswana musical videos.
- **Danger of Language Extinction:**
  The Sotho-Tswana languages face the risk of extinction if the number of speakers and singers in the language does not increase. A lack of technological tools supporting these languages [11] can demotivate artists from creating content in Sotho-Tswana languages, exacerbating the threat of extinction. Providing robust technological support can encourage artists to produce more content in these languages, thereby aiding in their preservation.

Addressing these challenges is crucial for the preser-

vation and advancement of Sotho-Tswana languages. By developing a multimodal language identification tool, this study aims to create a more inclusive and effective means of identifying and recommending Sotho-Tswana musical videos. This advancement not only enhances the user experience on social media platforms but also contributes to the cultural and linguistic vitality of Sotho-Tswana communities.

1.2. Aim and Objectives

The study aims to utilize a multimodal approach—encompassing audio, visual, and lyric modalities—along with deep learning and language models to identify the language used in Sotho-Tswana music videos. The specific objectives are outlined as follows:

- **Data Acquisition:** Download and access the Sotho-Tswana music dataset as specified in the data accessibility section of the published dataset [12].
- **Data Preprocessing:** Preprocess the multimodal dataset by removing blank instances and performing other preprocessing tasks.
- **Model Training and Validation:** Train and validate VGG16, a deep convolutional neural network model, for the visual modality, along with an artificial neural network (ANN) model for the audio modality and the Bidirectional Encoder Representation from Transformer (BERT) model for the lyrics/textual modality.
- **Performance Evaluation:** Evaluate the performance of the training and validation phases.
- **Result Fusion:** Use a late fusion method to combine the results of the training and validation for the audio, visual, and textual modalities.
- **Model Testing:** Test the trained models with sample audio, visual, and lyrics segments from video clips.

The rest of the paper is structured as follows:

- **Section 2:** Materials and Methods.
- **Section 3:** Presents the results of the visual, audio, textual, and multimodalities.
- **Section 4:** Provides a discussion of the findings.
- **Section 5:** Conclusion.

By systematically addressing these objectives, the study aims to enhance the identification of languages in Sotho-Tswana musical videos, thereby contributing to the preservation and promotion of these languages.

2. Materials and Methods

2.1. Language Identification in Speech

Language identification in music originated from language identification in speech, a critical task in multilingual speech processing. This process involves identifying the spoken language in audio data, which is a foundational step in developing multilingual systems such as multilingual speech recognition, information retrieval, and spoken language translation [2]. Furthermore, language identification in speech is crucial for making automatic speech recognition more inclusive [13].

From a linguistic perspective, spoken languages differ in traits such as phonology, prosody, vocabulary, and grammar [2]. Humans typically use one or more of these traits to identify a spoken language in an utterance. Similarly, language identification research has focused on traits like phonology and prosody, developing computationally efficient and reliable methods for language identification in speech. These methods can extract acoustic signals without requiring extensive language-specific knowledge [2].

Different machine-learning models have been employed to improve the accuracy of language identification. For instance, Bhola et al. [14] used a multi-layer perceptron with a sequential model, while Farris and Khudanpur [15] used Random Forest and Decision Tree techniques. A novel approach combining self-supervised representation learning with language labels was utilized by Vashishth et al. [16] for language identification in speech.

Additional methods include tokenization combined with phonotactic analysis [17] and Gaussian Mixture Models (GMM) with Shifted Delta Cepstral (SDC) coefficients [18]. Sugiyama [19] employed two algorithms for automatic language recognition in speech: one based on Vector Quantization and the other on a single universal Vector Quantization technique.

Different audio-based feature extraction techniques have been used for language identification in speech, they include the following:

- Mel Frequency Cepstral Coefficients (MFCC): Used by Bhola [14].
- Combined MFCC and GFCC Feature-Based BLSTM: Utilized by Adeeba and Hussain [20].
- Multi-Scale Feature Extraction Method (SE-Rev2Net-

CBAM-BILSTM): Employed by Aysa et al.[21] to achieve high accuracy.

- Linear Chirplet Transform (LCT): Demonstrated superiority over traditional methods like MFCC and Fourier Transforms[22].
- Bag of Words and Content-Based Audio-Visual Features: Used for language identification in speech[17].

Furthermore, multiple features integrated at the frame, pooling, and segment levels have been used for speaker and language identification tasks[23]. The study utilized acoustic features such as MFCC-FBank and MFCC-PLP for these audio classification tasks.

Language identification in speech typically requires a multilingual dataset to identify a specific language within a diverse set of languages used in audio data. Examples of such datasets include NIST LRE, BABEL, and KARAKA[24].

2.2. Language Identification in Music

Language identification in music has emerged as an important task in music information retrieval, offering several benefits such as:

- **Organizing Multilingual Music:** Categorizing music according to language[25].
- **Personalized Recommendations:** Enhancing user experience with tailored music suggestions[26].
- **Distinguishing Similar Tunes:** Identifying songs with the same melody but different lyrics[2].

Additionally, language identification in music aids melody-based music information retrieval in handling multilingual music documents[2]. However, the level of research on language identification in music remains relatively low compared to speech, with a predominant focus on audio music[2, 25].

Different approaches can be used for music identification in music, they include the following:

(1) **Phonetic Approach:**

This approach characterizes language-specific events and their distribution using acoustic features such as:

- **Mel Frequency Cepstral Coefficients (MFCC):** Used by Schwenninger et al.[27].
- **Stabilized Auditory Images (SAI):** Employed by Chandrasekhar et al.[25].
- **Temporal Patterns:** Utilized by Kruspe et al.[28].

(2) **Phonotactic Approach:**

The phonotactic approach identifies phonemes in an audio file and examines their combinations and sequences to identify the language, as each language has distinctive phoneme patterns[29]. Renault et al.[9] used this approach by combining an acoustic model for phoneme estimation with a language classifier to determine the language of an audio music track.

(3) **Acoustic Features and Machine Learning Models:**

Given that singing is a specialized form of speech, similar acoustic features and machine learning models used for language identification in speech have been adapted for music[26].

Finally, despite the relatively low level of research on language identification in music compared to speech, various approaches have been developed, each leveraging different acoustic features and machine-learning models.

2.3. Multimodal Approach to Language Identification in Music

The multimodal approach to language identification in music involves using combinations of different modalities, such as audio, lyrics/text, and visual data, to identify the singing language effectively. This approach has significantly improved the accuracy of language identification in music[26]. While most studies have focused primarily on audio, the availability of language information in other modalities has led to the development of a multimodal approach to language identification in music.

Like language identification in speech, multimodal language identification in music requires multilingual datasets. Carrasquillo et al.[18] used a dataset comprising musical videos in 25 different languages, incorporating audio and visual features such as audio spectrograms, Mel-Frequency Cepstral Coefficients (MFCC), stabilized auditory images, global visual data, and motion cuboids. Choi and ByteDance[6] used the Music4All dataset, which includes language labels such as English, Hindi, Slovak, Bulgarian, Hebrew, Portuguese, Spanish, Korean, French, Japanese, German, Polish, and Italian.

Different studies have employed various multimodal features to identify the singing language in music:

- Renault et al.[9] used multiple audio-based low-level features.

- Chandrasekhar et al.[25] combined low-level audio features and visual features.
- Choi and ByteDance et al.[6] used a combination of low-level generic song features and simple metadata fields such as album title and artist name.
- Lee and Coviello[26] employed richer audio representations, including vocal embeddings combined with high-level representations of timbral content.

Different methodologies have been employed in multimodal language identification in music:

- Carrasquillo et al.[18] used an early fusion method to combine various features into a single feature vector and trained a Support Vector Machine (SVM).
- Vashishth et al.[16] used a multilingual dataset to train the IDResNet architecture.
- Bhanja et al.[8] trained a combination of deep convolutional neural networks, Multi-Layer Perceptrons (MLPs), and modality dropout techniques.

The results obtained from all reviewed studies on the multimodal approach to language identification in music demonstrated that using a multimodal approach improved performance compared to using a single modality, such as audio alone[26].

Therefore, the integration of multiple modalities in language identification for music has proven to be more effective than relying on a single modality. By leveraging the combined information from audio, lyrics/text, and visual data, researchers can achieve higher accuracy in identifying the singing language, thereby enhancing the overall performance of music information retrieval systems.

2.4. Methodology

This study's methodology for a multimodal approach to language identification in Sotho-Tswana musical videos involves several key components. These components include:

(1) Data Access Techniques: Methods for accessing the dataset from a comprehensive multimodal music dataset of Sotho-Tswana musical videos.

(2) Fusion Methods: Techniques for integrating multiple modalities, including textual, audio, and visual data, to enhance the accuracy and robustness of language identification, Shivakumar et al.[30].

(3) Deep Learning and Language Model Architectures: The implementation of advanced deep learning frameworks and language models tailored for training on integrated multimodal data.

By combining these elements, this approach aims to improve the precision and efficiency of language identification in multimedia content, especially within the context of Sotho-Tswana musical videos.

2.4.1. Dataset Description

The multimodal dataset utilized to train models for various modalities is the multimodal dataset of Sotho-Tswana musical videos, which is accessible via Mendeley Data[31]. This dataset was curated specifically for diverse multimodal music information retrieval tasks, including a multimodal approach to language identification. Though a comprehensive guide on how to reproduce the dataset is available in[12], however, the visual and audio modalities of the downloaded Sotho-Tswana musical videos were separated using appropriate Python codes, provided by the dataset creators, while the textual/lyrics modality of the downloaded musical videos were obtained with the help of native speakers of Sotho-Tswana languages, who listened and watched the musical videos, and translated the audio modality into text. The native speakers of Sotho-Tswana languages also acted as Annotators, who labelled the musical videos, based on the following metadata of the dataset, Language, Lyrics, Genre, and Sentiment.

Comprising audio, textual (lyrics), and visual modalities of Sotho-Tswana musical videos, this dataset underwent several stages of curation, as outlined by[32]. These stages included data acquisition, data pre-processing, and data annotation. The recommended steps for downloading and accessing the dataset were followed as per[12]. The videos were downloaded and segmented into fifteen-second video segments. The distribution of the video segments according to different Sotho-Tswana languages is shown in **Figure 1**.

2.4.2. Fusion Methods

Fusion methods delineate how the different modalities, such as textual (lyrics), audio, and visual modalities will be integrated. Among these, the two most prevalent fusion methods are the feature-level (early) fusion method and the decision-level (late) fusion method. However, due to its simplicity and ease of use advantage, the decision-level (late) fusion method was recommended for the multimodal dataset of Sotho-Tswana musical videos, which was employed in

this study. Both fusion methods will be succinctly elucidated in the subsequent sub-section.

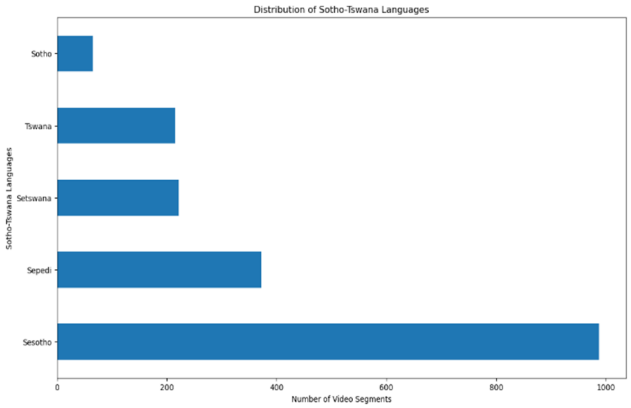

Figure 1. Dataset distribution of Sotho-Tswana musical video segments based on Sotho-Tswana languages.

Feature Level (Early) Fusion Method

In the feature-level or early fusion technique, the features of each modality under consideration are extracted and combined into one feature vector before training. One challenge associated with this fusion method is the complexity of combining features from different modalities with varying dimensions. Due to this difficulty, this fusion technique is not commonly employed compared to the decision-level or late fusion technique. **Figure 2** illustrates the feature-level or early fusion method.

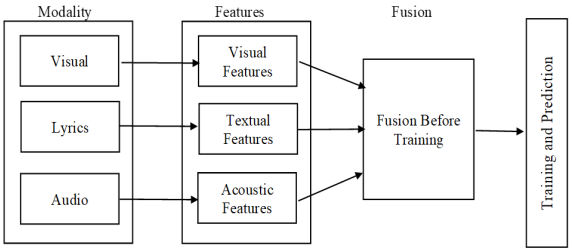

Figure 2. Feature level (early) fusion method.

Decision Level (Late) Fusion Method

In the decision-level or late fusion method, the features of each modality under consideration are extracted and trained independently using specific models, and the results of the training are combined to obtain the final result. This study employed this fusion technique to determine the language used in the textual, visual, and audio modalities of Sotho-Tswana musical videos, and afterward, combine them. **Figure 3** illustrates the decision-level or late fusion method.

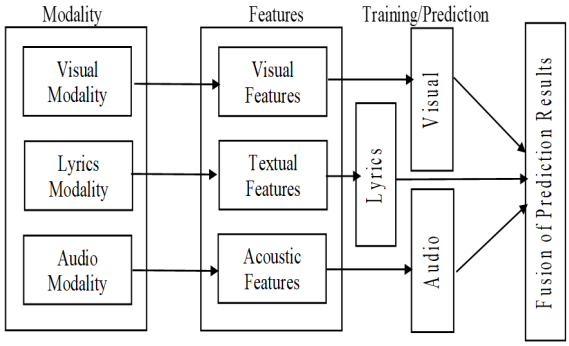

Figure 3. Decision level (late) fusion method.

Following the independent training of the three modalities, late fusion is applied to combine the results, obtaining the singing language, by averaging the outcomes of the three modalities. For instance, with 5 different languages, each assigned a distinct score from 1 to 5, after training and validation, if the results of a sample video segment for the three modalities are 1, 2, and 3 respectively, the average score would be 2. The language corresponding to the assigned value of 2 would then be the result of the multimodal prediction. In cases where the average is a fraction, it is approximated to the nearest whole number.

2.4.3. Deep Learning and Language Models

While various deep learning models can be trained for the visuals and audio modalities [32], we opted to utilize the VGG16 model for the visual modality. VGG16 is a deep convolutional neural network model selected for its proficiency in detecting and classifying images across diverse categories with high accuracy.

Conversely, for the audio modality, we have chosen the Artificial Neural Network (ANN) model, implemented through the Keras Sequential API. This model will be employed for the audio modality of singing language identification. On the other hand, we have chosen a language model, BERT (Bidirectional Encoder Representation from Transformer) to train the textual modality.

Architecture of VGG16 and ANN Models

VGG16, a deep convolutional neural network renowned for its efficacy in image recognition tasks, boasts a formidable architecture comprising 16 trainable layers. This architecture encompasses 13 convolutional layers, strategically interspersed with 5 max pooling layers, and culminating in 3 dense layers, thereby constituting a total of 21 layers. Detailed schematics depicting the network's

structural intricacies are provided in [33] for reference and visualization.

The architecture of VGG16 dictates an input tensor size of 224 x 224 x 3, setting the stage for its subsequent convolutional operations. Convolutional layers within VGG16 are meticulously designed with varying filter capacities to extract hierarchical features. Specifically, convolutional layer 1 is endowed with 64 filters, layer 2 escalates this to 128 filters, and layer 3 further amplifies feature extraction with 256 filters. Layers 4 and 5 of the convolutional stack exhibit even greater prowess, each boasting 512 filters. Following this convolutional cascade, the network transitions into a trio of fully connected layers, each pivotal in refining extracted features for classification. The initial two fully connected layers are characterized by 4096 channels each, while the final layer assumes the responsibility of mapping these features into 1000 distinct classes, as elucidated in [33].

Conversely, in the realm of audio modality analysis, an artificial neural network (ANN) model, constructed utilizing the Keras API, assumes prominence. This model is tailored to discern the class or polarity of audio clips, embodying a streamlined architecture comprising three dense layers. The first dense layer, housing 100 neurons, interfaces with an input shape of 40 and employs the Rectified Linear Unit (ReLU) activation function to foster non-linearity. Subsequent dense layers, each housing 200 neurons, integrate a dropout rate of 0.5 to mitigate overfitting concerns, while still leveraging the ReLU activation function to facilitate feature extraction and learning.

Architecture of BERT Language Model

BERT, short for Bidirectional Encoder Representation from Transformer, stands as a pivotal advancement in neural network technology tailored for natural language processing tasks, notably including textual/lyrics-based singing language identification. Leveraging a transformer-based architecture, BERT emerges as a pre-trained language model, honed through exposure to extensive volumes of unlabeled text data drawn from Wikipedia.

Central to its design lies a multi-layer bidirectional transformer encoder framework, manifesting in two primary variants: BERTbase and BERTlarge. BERTbase encompasses a 12-layer Encoder stack, whereas BERTlarge extends this architecture with 24 layers. Both variants integrate expansive feed-forward networks, featuring 768 hidden layers for BERTbase and 1024 for BERTlarge, complemented by 12 and 16 attention heads, respectively. Notably, BERTbase encapsulates 110 million parameters, while BERTlarge amplifies this complexity with 340 million parameters.

An integral element within the BERT architecture is the CLS token, serving as a classification anchor preceding a sequence of input words. This input sequence traverses through the layers of BERT, undergoing self-attention mechanisms and traversing feedforward networks before progressing to subsequent encoders. Ultimately, the model yields a vector output with a hidden size of 768 for BERTbase, thus facilitating nuanced representations of textual data.

2.5. Experimental Design for Model Training

The dataset for multimodal music information retrieval of Sotho-Tswana music videos, which was downloaded from Mendeley Data was used to generate about 16,983 instances of video images/frames dataset, which was utilized to train the visual modality. An equivalent number of instances of data was employed to train the audio and textual modalities. Part of the preprocessing conducted prior to training involved ensuring that all the relevant metadata were not blank.

Before commencing training, the datasets for textual, audio, and visual modalities were divided, with 80% allocated for the training set and 20% for the testing set. Subsequently, the models for audio and visual modalities were trained independently. Similarly, the model for the textual modality (lyrics) underwent independent training. Following the training of the models for the three modalities, the results were amalgamated using the late fusion method by obtaining the average of the three modalities.

The independent training of the models for the audio and visual modalities proceeded as follows: Initially, all relevant libraries were loaded, and subsequently, the two CSV files for the three modalities, AudioSegments.csv and VideoSegment.csv, were read using Pandas. The visual modality was then trained by loading all images/frames via the image_load package in Keras, and storing them in an array. VGG16 was employed to extract features from the training and testing images using the pixel values of the images. Train accuracy, validation accuracy, loss, and validation loss were computed, followed by making predictions with test data.

Similarly, for the audio modality, librosa was utilized to

extract the MFCC features for each of the segmented audio clips. The dataset was then split into a training set and a test set, utilizing a ratio of 0.8 for the train set and 0.2 for the test set. The architecture of the ANN model was defined and employed to fit the train set and test set. Post-fitting, accuracy, validation accuracy, loss, and validation loss were measured, and finally, predictions were made with test data.

Furthermore, for the training of the model for the textual (lyrics) modality, the training and test datasets were initially split using the ratio 0.8 for training and 0.2 for the test set. Subsequently, they were read using pandas read_csv. The language labels for the train and test sets were mapped with numbers using a Python dictionary. They were then converted into categorical columns using to_categorical. The lyrics for the training and testing set, which constitute the input texts, were converted into BERT input data format using AutoTokenizer, Bert-base-cased. This was achieved by passing all necessary parameters to the Tokenizer, including add_special_tokens = True, max_length = 70, truncation = True, padding = True, return_tensors = 'tf', return_token_type_ids = False, return_attention_mask = True, and verbose = True. After tokenization, the tokenizer returned a dictionary containing input_ids and attention_masks.

The model was designed and built using the Keras functional API. Since the BERT model employs three parameters—input_ids, attention_masks, and token_type_ids—these were passed to the BERT model, except token_type_ids, which is unnecessary for sentiment analysis. The input layers of the BERT models are the input_ids and attention_masks, while the embeddings contain the hidden states of the BERT layer. CNN layers, which yield the output, were constructed using the hidden states of the BERT. For building the CNN layers atop the BERT model, BERT hidden forms were utilized, where BERT[0] represents the last hidden state and BERT[1] is the pooler output. The model compilation was kept simple, with the following hyperparameters: loss = "categorical_crossentropy", optimizer = 'adam', epoch = 20, batch size = 128.

3. Results

After the training and validation, with appropriate hyper-parameters for optimal performance, the following performance metrics were recorded for twenty different epochs, loss, accuracy, val_loss, and val_accuracy. T**able 1, Table 2, Table 3,** and **Table 4** show the results of the four-performance metrics, for the visual, audio, lyrics, and multimodality, respectively. Furthermore, **Figure 4**, **Figure 5**, **Figure 6**, and **Figure 7** show the visualizations of the performance metrics, for the visual, audio, lyrics, and multimodality, respectively.

Table 1. Performance metrics for the visual modality.

	Visual Modality			
Epoch	Loss	Accuracy	Test_Loss	Test_Accuracy
1	0.5918	0.7724	0.0121	0.9953
2	0.0349	0.9909	4.70E-03	0.9985
3	0.0246	0.9936	4.60E-03	0.9988
4	2.69E-02	0.9924	2.70E-03	0.9994
5	1.74E-02	0.9957	4.00E-03	0.9988
6	1.69E-02	0.9948	3.40E-03	0.9994
7	1.50E-02	0.9957	4.10E-03	0.9988
8	1.98E-02	0.9943	1.05E-02	0.9965
9	1.73E-02	0.9952	1.70E-03	0.9994
10	1.11E-02	0.9969	2.00E-03	0.9994
11	1.04E-02	0.997	3.20E-03	0.9991
12	7.80E-03	0.9977	2.30E-03	0.9991
13	1.03E-02	0.9968	1.90E-03	0.9991
14	1.20E-02	0.9968	1.50E-03	0.9994
15	8.80E-03	0.9974	1.10E-03	0.9997
16	8.60E-03	0.998	1.40E-03	0.9997
17	9.60E-03	0.997	2.20E-03	0.9994
18	6.20E-03	0.9983	3.80E-03	0.9991
19	5.10E-03	0.9986	2.60E-03	0.9991
20	6.00E-03	0.9982	2.20E-03	0.9991
Average	0.043025	0.984885	0.0036	0.998855

The results of **Table 1** were obtained by tunning the hyper-parameter epoch, from 1 to 20, and recording the results obtained by the VGG16, which was used to train the visual modality.

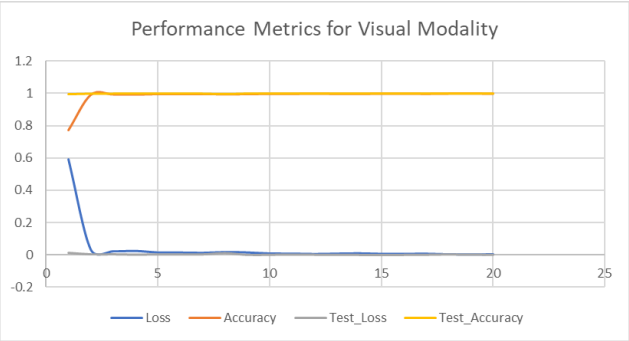

Figure 4. Visualization of the performance metrics for the visual modality.

Figure 4 was obtained by plotting the various perfor-

mance metrics, Loss, Accuracy, Test_Loss, and Test_Accuracy against the hyper-parameter, epoch, for the visual modality.

Table 2. Performance metrics for the audio modality.

	Audio Modality			
Epoch	Loss	Accuracy	Test_Loss	Test_Accuracy
1	0.1058	0.987	2.90E-03	1
2	0.1072	0.9665	3.30E-03	1
3	0.0926	0.973	2.40E-03	1
4	0.0743	0.9784	1.10E-03	1
5	0.0703	0.9796	5.88E-04	1
6	0.0614	0.9825	3.35E-04	1
7	0.0595	0.9836	8.42E-04	1
8	0.0537	0.9842	4.17E-04	1
9	0.0537	0.9858	3.74E-04	1
10	0.0489	0.9862	2.24E-04	1
11	0.0464	0.9876	9.44E-04	1
12	0.0416	0.989	1.73E-04	1
13	0.0399	0.988	2.34E-04	1
14	0.0413	0.9887	1.44E-04	1
15	0.0446	0.9882	2.25E-04	1
16	0.0463	0.9878	1.40E-04	1
17	0.0457	0.988	4.28E-04	1
18	0.0991	0.9905	6.45E-04	1
19	0.0326	0.9915	1.54E-04	1
20	0.0339	0.9916	9.21E-05	1
Average	0.05994	0.984885	0.00078297	1

The results of **Table 2** were obtained by tunning the hyper-parameter epoch, from 1 to 20, and recording the results obtained by the ANN model, which was used to train the audio modality modality.

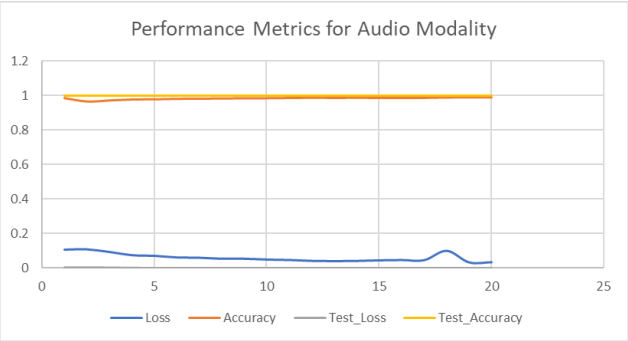

Figure 5. Visualization of the performance metrics for the audio modality.

Figure 5 was obtained by plotting the various performance metrics, Loss, Accuracy, Test_Loss, and Test_Accuracy against the hyper-parameter, epoch, for the visual modality.

Table 3. Performance metrics for the textual/lyrics modality.

	Textual/Lyrics Modality			
Epoch	Loss	Accuracy	Test_Loss	Test_Accuracy
1	1.4964	0.4008	1.4923	0.4026
2	1.454	0.4154	1.4885	0.4026
3	1.4516	0.4154	1.4881	0.4026
4	1.4524	0.4156	1.4922	0.4026
5	1.4504	0.4157	1.4766	0.4026
6	1.4508	0.4157	1.478	0.4026
7	1.4547	0.4152	1.5196	0.4026
8	1.4505	0.4157	1.4749	0.4026
9	1.4499	0.4157	1.4687	0.4026
10	1.4511	0.4157	1.4727	0.4026
11	1.4503	0.4157	1.4754	0.4026
12	1.4491	0.4157	1.4688	0.4026
13	1.4586	0.4124	2.0216	0.2045
14	1.4537	0.4158	1.9208	0.2045
15	1.4512	0.4157	1.9597	0.2045
16	1.451	0.4157	2.019	0.2045
17	1.4511	0.4157	1.9839	0.2045
18	1.4509	0.4157	1.9941	0.2045
19	1.4508	0.4157	2.0342	0.2045
20	1.4508	0.4157	2.1086	0.2045
Average	1.453965	0.414735	1.691885	0.32336

The results of **Table 3** were obtained by tunning the hyper-parameter epoch, from 1 to 20, and recording the results obtained by the language model, BERT model, which was used to train the textual/lyrics modality modality.

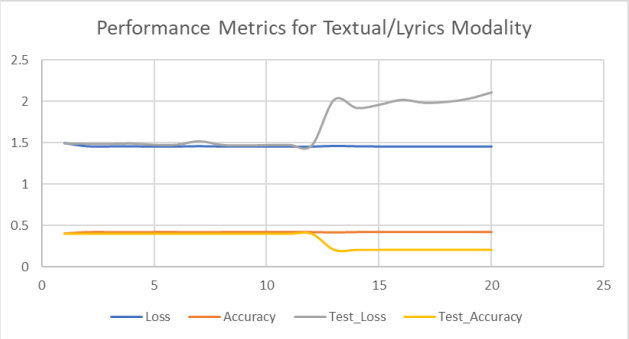

Figure 6. Visualization of the performance metrics for the lyrics/textual modality.

The results of **Table 4** were obtained by averaging the results of the performance metrics for the three modalities, for each of the tunned hyper-parameters, epoch.

Figure 6 was obtained by plotting the various performance metrics, Loss, Accuracy, Test_Loss, and Test_Accuracy against the hyper-parameter, epoch, for the visual modality.

Table 4. Performance metrics for the multimodality.

	Multimodality (Audio, Visual and Lyrics)			
Epoch	Loss	Accuracy	Val_Loss	Val_Accuracy
1	0.7313333	0.7200667	0.502433	0.7993
2	0.5320333	0.7909333	0.498833	0.800366667
3	0.5229333	0.794	0.498367	0.800466667
4	0.5178667	0.7954667	0.498667	0.800666667
5	0.5127	0.797	0.493729	0.800466667
6	0.5097	0.7976667	0.493912	0.800666667
7	0.5097333	0.7981667	0.508181	0.800466667
8	0.508	0.7980667	0.495272	0.7997
9	0.5069667	0.7989	0.490258	0.800666667
10	0.5037	0.7996	0.491641	0.800666667
11	0.5023667	0.8001	0.493181	0.800566667
12	0.4995	0.8008	0.490424	0.800566667
13	0.5029333	0.7990667	0.674578	0.734533333
14	0.5023333	0.8004333	0.640815	0.734633333
15	0.5015333	0.8004333	0.653675	0.734733333
16	0.5019667	0.8005	0.673513	0.734733333
17	0.5021333	0.8002333	0.662176	0.734633333
18	0.5187333	0.8015	0.666182	0.734533333
19	0.4961667	0.8019333	0.678985	0.734533333
20	0.4969	0.8018333	0.703631	0.734533333
Average	0.5189767	0.794835	0.565423	0.774071667

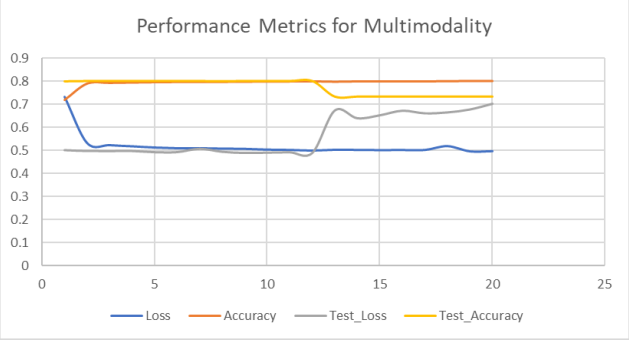

Figure 7. Visualization of the performance metrics for the multimodality.

Figure 7 was obtained by plotting the various performance metrics, Loss, Accuracy, Test_Loss, and Test_Accuracy against the hyper-parameter, epoch, for the multimodality.

4. Discussion

The results for the performance metrics, loss, accuracy, test_loss, and test_accuracy range from 0 to 1, however, for best performance results, the loss and test_loss metrics are expected to be close to 0, while the accuracy and test_accuracy metrics are expected to be close to 1. From **Table 1**, **Table 2**, and **Table 3**, the averages of the various performance metrics for visual modality and audio modality perform better than the averages of the various performance metrics for the lyric's modality. Furthermore, the results in **Table 1**, **Table 2**, and **Table 3**, together with **Figure 1**, **Figure 2**, and **Figure 3** show the consistency of the results of the various performance metrics. For all the values of the hyper-parameter, epoch, the loss and test_loss metrics remains low, while Accuracy and test_Accuracy metrics remain high. The tunning of the hyper-parameter, epoch on the results shows that the results for the visual modality perform better than the results for the audio and lyrics modalities, for all the performance metrics. However, the results of the multimodality, which is shown in **Table 4**, were obtained by taking the average of the corresponding results for the visual, audio, and lyrics modalities. This result shows that the averages of the various performance metrics for the multimodality are better than the average results of the various performance metrics for the lyrics modality. This result aligns with the fact that music information is embedded in the combinations of the various modalities, rather than in one modality alone, like lyrics, which is consistent with similar results obtained in literature [26].

5. Conclusions

This paper has employed a multimodal approach to identify the singing language in Sotho-Tswana musical videos. It achieved this by utilizing three different modalities embedded in these videos: audio, visual, and textual (lyrics). The multimodal dataset used is the Sotho-Tswana musical videos available in the Mendeley public repository. The visual modality of the dataset was used to train the VGG16 deep convolutional neural network, the audio modality of the dataset was used to train the Artificial Neural Network, while the textual/lyrics modality of the dataset was used to train Bidirectional Encoder Representation from Transformer, (BERT). The results of the independent training were combined using the decision level (late) fusion method by applying the averaging technique to the results of the training of the three modalities. The results indicate that combining the three modalities is more effective than relying on a single modality, such as text or lyrics.

Author Contributions

Conceptualization, O.E.O.; methodology, O.E.O.; software, O.E.O.; validation, O.E.O.; formal analysis, O.E.O.; investigation, O.E.O.; resources, M.P.; data curation, O.E.O.; writing—original draft preparation, O.E.O.; writing—review and editing, O.E.O.; visualization, O.E.O.; supervision, M.P.; project administration, M.P.; funding acquisition, M.P. All authors have read and agreed to the published version of the manuscript.

Funding

This work received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The dataset used in this paper is the multimodal dataset of the Sotho-Tswana musical video, which is available in a public repository, Mendeley Data. The data reference of this dataset follows below:

Osondu Oguike and Mpho Primus, "A Dataset for Multimodal Music Information Retrieval of Sotho-Tswana Music Videos," Mendeley Data, V1, 2024. doi: 10.17632/7jmgfk4fd9.1

Acknowledgments

O. E. Oguike acknowledges the use of ChatGPT in this study, which was used for copy editing to improve language readability.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Mehrabani, M., Hansen, J.H.L., 2011. Language Identification for Singing. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 22 May-27 May 2011; Prague, Czech Republic. pp. 4408–4411.

[2] Tsai, W.H., Wang, H.M., 2004. Towards Automatic Identification of Singing Language in Popular Music Recordings. In International Society for Music Information Retrieval, October 2004, Available from: https://homepage.iis.sinica.edu.tw/papers/whm/1384-F.pdf

[3] He, H., Jin, J., Xiong, Y., et al., 2008. Language Feature Mining for Music Emotion Classification via Supervised Learning from Lyrics. Proceedings of Advances in the 3rd International Symposium on Computation and Intelligence (ISICA 2008); 19–21 December 2008; Wuhan, China. pp. 426–435.

[4] Chen, Z., Liu, C., 2021. Music Audio Sentiment Classification Based on CNN-BiLSTM and Attention Model. IEEE 4th International Conference on Robotics, Control and Automation Engineering; 04 November–06 November 2021; Wuhan, China. pp. 156–160.

[5] Pasrija, S., Sahu, S., Meena, S., 2023. Audio Based Music Genre Classification using Convolutional Neural Networks Sequential Model. IEEE 8th International Conference for Convergence in Technology (I2CT); 7–9 April 2023; Pune, India. pp. 156–160.

[6] Choi, K., ByteDance, Y.W., 2021. Listen, Read and Identify: Multimodal Singing Language Identification of Music. Proceeding of the 22nd International Society for Music Information Retrieval Conference; arXiv preprint arXiv:2103.01893. pp. 1–7.

[7] Mukherjee, H., Dhar, A., Obaidullah, S.M., et al., 2021. Identifying Language from Songs. Multimedia Tools and Applications. 80(28), 35319–35339.

[8] Bhanja, C.C., Laskar, M.A., Laskar, R.H., et al., 2022. Deep Neural Network Based Two-Stage Indian Language Identification System Using Glottal Closure Instants as Anchor Points. Journal of King Saud University – Computer and Information Sciences. 34(4), 1439–1454.

[9] Renault, L., Vaglio, A., Hennequin, R., 2021. Singing Language Identification Using Deep Phonotactic Approach. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 6–11 June 2021; Toronto, ON, Canada. pp. 271–275.

[10] Chavula, C., Suleman, H., 2021. Ranking by Language Similarity for Resource Scarce Southern Bantu Languages. Proceedings of the 2021 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '21), 11 July 2021; Virtual Event, Canada. ACM, New York, USA. pp. 1–11.

[11] Liu, Z., Richardson, C., Hatcher, R., Jr., et al., 2022. Not Always About You: Prioritizing Community Needs When Developing Endangered Language Technology. arXiv preprint arXiv:2204.05541.

[12] Oguike, O., Primus, M., 2024. A Dataset for Multimodal Music Information Retrieval of Sotho-Tswana Musical Videos. Data in Brief. 55, 110672.

[13] Sharimbaev, B., Kadyrov, S., 2023. Automatic Language Identification from Audio Signals Using LSTM-RNN. 17th International Conference on Electronics Computer and Computation (ICECCO); 1–2 June 2023; Kaskelen, Kazakhstan. pp. 1-5.

[14] Bhola, A., Reddy, K.N., Kumar, M.J., 2023. Language Identification using Multi-Layer Perceptron. International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES); 28–30 April 2023; Greater Noida, India; pp. 1018–1022.

[15] Farris, D., White, C., Khudanpur, S., 2008. Sample Selection for Automatic Language Identification. IEEE International Conference on Acoustics, Speech and Signal Processing; 31 March–4 April 2008; Las Vegas, NV, USA. pp. 4225–4228.

[16] Vashishth, S., Bharadwaj, S., Ganapathy, S., et al., 2023. Label Aware Speech Representation Learning for Language Identification. Proceedings of Interspeech; arXiv preprint arXiv:2306.04374. pp. 5351–5355.

[17] Zissman, M.A., 1996. Comparison of Four Approaches to Automatic Language Identification of Telephone Speech. IEEE Transactions on Speech and Audio Processing. 4(1), 31–44.

[18] Carrasquillo, P.A.T., Singer, E., Kohler, M.A., et al., 2002. Approaches to Language Identification Using Gaussian Mixture Models and Shifted Delta Cepstral Features. Proceeding of International Conference on Spoken Language Processing; 16 September 2002; Denver, USA. pp. 89–92.

[19] Sugiyama, M., 1991. Automatic Language Recognition Using Acoustic Features. Proceedings ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing; 14–17 April 1991; Toronto, ON, Canada. pp. 813–816.

[20] Adeeba, F., Hussain, S., 2019. Native Language Identification in Short Utterances Using Bidirectional Long Short-Term Memory Network. IEEE Access. 7, 17098–17110.

[21] Aysa, Z., Ablimit, M., Hamdulla, A., 2023. Multi-Scale Feature Learning for Language Identification of Overlapped Speech. Applied Science. 13(7), 4235. DOI: https://doi.org/10.3390/app13074235.

[22] Do, H.D., Chau, D.T., Tran, S.T., 2023. Speech Feature Extraction Using Linear Chirplet Transform and Its Applications. Journal of Information and Telecommunication. 7(3), 376–391. DOI: https://doi.org/10.1080/24751839.2023.2207267.

[23] Li, Z., Zhao, M., Li, J., et al., 2020. On the Usage of Multi-Feature Integration for Speaker Verification and Language Identification. Proceedings of Interspeech 2020; 25–29 October 2020; Shanghai, China. pp 457–461.

[24] Tan, Z., Wang, D., Chen, Y., et al., 2018. Phonetic Temporal Neural Model for Language Identification. IEEE/ACM Transaction on Audio, Speech and Language Processing. 26(1), 134–141.

[25] Chandrasekhar, V., Sargin, M.E., Ross, D.A., 2011. Automatic Language Identification in Music Videos with Low Level Audio and Visual Features. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 5724–5727.

[26] Lee, W.J., Coviello, E., 2022. A Multimodal Strategy for Singing Language Identification. Proceedings of Interspeech 2022; 18–22 September 2022; Incheon, Korea. pp. 2243–2247.

[27] Schwenninger, J., Brueckner, R., Willett, D., et al., 2006. Language Identification in Vocal Music. Proceeding of International Society for Music Information Retrieval, (ISMIR 2006). pp. 377–379.

[28] Kruspe, A.M., Abesser, J., Dittma. C., 2014. A GMM Approach to Singing Language Identification. AES 53RD International Conference; 27–29 January 2014; London, UK. pp. 1–9.

[29] Li, H., Ma, B., Lee, K.A., 2013. Spoken Language Recognition: From Fundamentals to Practice. Proceeding of the IEEE. 101(5), 1136–1159.

[30] Shivakumar, P.G., Chakravarthula, S.N., Georgiou, P.G., 2016. Multimodal Fusion of Multirate Acoustic, Prosodic, and Lexical Speaker Characteristics for Native Language Identification. Proceedings of Interspeech September 2016. pp. 2408–2412.

[31] Oguike, O., Primus, M., 2024. A Dataset for Multimodal Music Information Retrieval of Sotho-Tswana Music Videos. Data in Brief. 55, 110672.

[32] Gandhi, A., Adhvaryu, K., Poria, S., et al., 2023. Multimodal Sentiment Analysis: A Systematic Review of History, Datasets, Multimodal Fusion Methods, Applications, Challenges, and Future Directions. Information Fusion. 91, 424–444.

[33] Sarker, S., Tushar, S.N.B., Chen, H., 2023. High Accuracy Keyway Angle Identification Using VGG16-Based Learning Method. Journal of Manufacturing Processes. 98, 223-233.