lang-identification-revised

by Ye Kyaw Thu

Automatic Language Identification on Audio Signals

Khaing Zar Mon
ME-IST-5
Department of Information Science
University of Technology (Yatanarpon Cyber City)
khaingzarmon@utycc.edu.mm

May Phyu Khin
ME-IST-7
Department of Information Science
University of Technology (Yatanarpon Cyber City)
mayphyukhin@utycc.edu.mm

Myo Mar Thinn
ME-IST-3
Department of Information Science
University of Technology (Yatanarpon Cyber City)
myomarthinn@utycc.edu.mm

Ei Thandar Phyu
ME-IST-2
Department of Information Science
University of Technology (Yatanarpon Cyber City)
eithandarphyu@utycc.edu.mm

Nang Aeindray Kyaw
ME-IST-6
Department of Information Science
University of Technology (Yatanarpon Cyber City)
nangaeindraykyaw@utycc.edu.mm

Abstract—The purpose of our study is to identify which language is spoken from a speech sample, based solely on the acoustic information conveyed by the speech signal. There are many characteristics of speech that could be used to identify languages. Languages are made up of different sounds that form phonemes, so it is possible to distinguish languages based on the acoustic features present in the speech signal. We decided to study language identification of four different languages, namely English, Chinese, Myanmar (Burmese) and Shan (Tai Long). In this study, three different approaches were used to carry out the experiments. The first one is language identification from Mel Frequency Cepstral Coefficients (MFCC) using Keras Library. In the second approach, we use waveforms of raw audio signals as input to a Residual Neural Networks (ResNets). For the last approach, we use spectrograms of our audio data as input to ResNets which in turn is trained for language identification. According to our experimental results, all the approaches give promising results. The first approach, language identification with MECC using Keras, and the last approach, identification using spectrograms, give 100% testing accuracy. We expect a more useful system can be developed as more data (with more language classes) comes available in the future.

Index Terms—Language Identification, Residual Neural Networks (ResNets), Mel Frequency Cepstral Coefficients (MFCC)

I. INTRODUCTION

Recently, voice assistants have become a staple in the flagship products of many big technology companies such as Google, Apple, Amazon, and Microsoft. One challenge for intelligent assistants, like Siri or the Google Assistant, is that the language that a speaker is using needs to be preset. To improve user experience on this and similar tasks such as automated speech detection or speech to text transcription, automatic language detection is a necessary first step. Moreover, emergency call routing can be one of the applications of language identification, where the response time of a fluent native operator might be critical. The main motivation is to study language identification from audio data using different Deep Learning approaches.

Finding a dataset of audio clips in various languages sufficiently large for training a network was an initial challenge for this task. Due to the difficulty to get open-source audio corpus for various languages, we only use audio corpus with four different languages, namely English, Chinese, Myanmar (Burmese) and Shan (Tai Long). In this project, we studied the implementation and evaluation of three different language identification models and compared the results. Firstly, we extracted Mel Frequency Cepstral Coefficients (MFCC) features from our audio data and trained with Keras Library to build a language identification model. We also studied to use Residual Neural Networks (ResNets) for language identification and built two models with ResNets. Waveforms of audio data are used as input to the second model and spectrograms of audio data are used as input of the third model.

The structure of this paper is organized as follows. Related works of language identification for different languages using audio data are presented in the upcoming section. Section III describes the methodologies used in the language identification experiments. Later, in Section IV, we present corpus preparation and data pre-processing together with feature extraction techniques used for the experiments. The experimental results along with some discussions are described in Section V. Eventually, in Section VI, we conclude our works showing promising results.

II. Related Work

Audio language identification systems are usually based on identity vector systems [4] and [9]. Spectorgram-based language classification are also performed on neural network models [1], [2] and [11]. Using deep neural network, language identification based on audio signals were also presented in [5], [8] and [10]. In [12], identification results using Long Short-Term Memory (LSTM) Recurrent Neural Network provided better performance than in identity vector systems. Based on related studies, these neural networks based systems are more convenient to language identification tasks because they achieved highest accuracy rates. Therefore, we performed language identification based on audio files and images (spectrograms and waveforms) for different four languages.

III. Methodology

In this section, we describe the methodologies used in language classification on audio features.

A. Audio Feature Extraction

Before an audio signal can be classified, the features within that signal first need to be extracted and analysed. These features represent the characteristics of the audio signal and will ultimately decide the class of that signal. The techniques employed during feature extraction can either involve the analysis of the actual waveform of the audio signal, or the analysis of the spectral representation of the audio signal. During the feature extraction stage there is reduction of data from the audio signal as sound data contains much redundancy [3]. This is done by breaking down the audio signal into successive short-time or short-term windows or frames, which are generally no larger than 100ms [6]. A set of features are then calculated for each frame, resulting in a feature vector [12].

B. Mel-frequency Cepstral Coefficients (MFCC)

MFCC is one of the most important method to extract a feature of an audio signal and is used majorly whenever working on audio signals. In practise, MFCC are the discrete cosine transform coefficients of the mel-scaled log-power spectrum. MFCC is a technique based on human hearing behavior that cannot recognize frequencies over 1 kHz. MFCC are based on the difference of frequencies that the human ear can distinguish. The signal is expressed in the MEL scale, this scale is based on the perception of the pitch in an equally spaced intervals judged by observers [7]. MFCCs have been widely used in speech recognition, speaker clustering and many other audio analysis applications.

C. Transfer Learning

Transfer learning is a method to leverage the generic knowledge acquired by a model trained on massive datasets to develop a more task-specific knowledge using only a limited amount of new data. While generic knowledge is used to detect low-level features in an image, such as edges and curves, task-specific knowledge enables the model to recognize higher-level features, such as the shape of a hand, type of spectrogram and wave images. Generic knowledge is formed in the lower layers of the CNN (near the input image), while specific knowledge grows in the upper layers, near the classifier (the output) [14].

D. ResNet-50

ResNet-50 is the Residual Network with 50 layers. Convolutional neural networks (CNN), a type of deep learning network with excellent performance in detecting features in images [15]. There are many pre-trained models like VGG, Inspection-V3, MobileNet,ResNet and so on. These pre-trained models saving us the time and expense of having to do the training from scratch. ResNet-50 is ideal for speed, accuracy and can load a pretrained version of the network trained on more than a million images from the ImageNet database. The pretrained network can classify images into 1,000 object categories, such as keyboard, mouse, pencil, and many animals. As a result, the network has learned rich feature representations for a wide range of images. The network has an image input size of 224-by-224 [16].

IV. Experiments

In this section, the experiment settings for the audio language identification are described. The experiment consisted of four main parts: preparation of audio data, audio featuring, classification of different languages based on audio files and classification of different languages based on images (spectrograms and waveforms).

A. Audio Data Preparation

For our experiments, audio clips were gathered in Burmese, English, Shan, and Chinese. We used topia - Microblog Translated Posts Parallel Corpus (Release V1.1 - 19/09/2013) for English and Chinese audio files. For Shan language, we prepared audio files by recording three native Shan speakers. Myanmar audio files were collected from weather broadcasts. Speakers had various accents and were of different genders. For each language, there were 1,000 audio files. All audio files were sampled at a rate of 16kHz with a mono channel. The audio files were saved as WAV files.

B. Audio Featuring

In this experiment, the audio features of audio signal are extracted from the audio files for audio based language classification. We used Librosa which is a python package used for analyzing and extracting features of an audio signal. For image based classification, the audio files were represented in the forms of images (spectrograms and waveforms). A spectrogram is a detailed view of audio which displays changes in the frequencies in a signal over time. A waveform is an image that represents an audio signal showing the changes in amplitude over a certain amount of time. The spectrograms of the audio files were
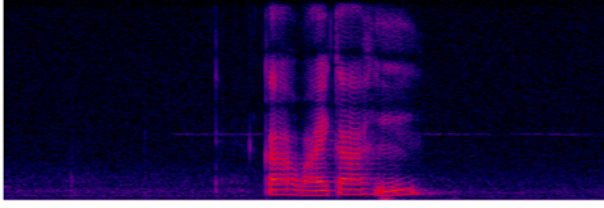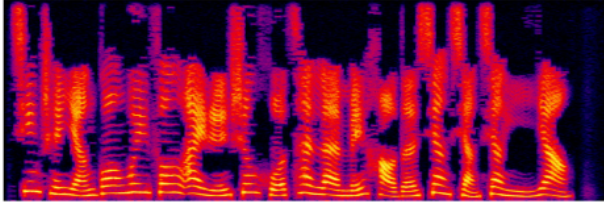
Fig. 1. Spectrogram of Shan Audio File

Fig. 3. Waveform of Burmese Audio File
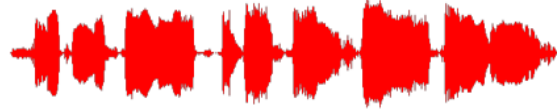

Fig. 2. Spectrogram of Chinese Audio File

Fig. 4. Waveform of English Audio File

extracted using the Sound eXchange (SoX, Swiss Army knife of sound processing programs) command line utility [10]. Using the FFmpeg cross-platform solution, the audio files were converted to the waveforms. Some spectrograms of Shan and Chinese audio files are described in Fig. 1 and Fig. 2 respectively. In Fig. 3 and Fig. 4, waveforms of Burmese and English languages are shown.

C. Classification of different langues based on audio files

In our experiments, the total audio files used for four languages are 4,000. The integrated development environment (IDE) used for the process is Google Colaboratory or Google Colab in short. Google Colab is a free Jupyter notebook environment that requires no setup and runs entirely in the cloud. With Google Colab, it is possible to write and execute code, save and share the analyses, and access powerful computing resources, all for free from the browser.

The audio features are extracted from the audio files. We used Librosa for analyzing and extracting features of an audio signal. Librosa is a python package used for audio analysis. After extracting features, the data is standardized by removing the mean and scaling to unit variance. Standardization is useful for negative-value results.

The language is predicted based on the Convolutional Neural Network (CNN). There are 4 layers in this network. The first one is an input layer therefore, input size is equal to the numbers of features used. Then there are 2 hidden layers and the last layer is an output layer. The output layer has 4 neurons as we are classifying into 4 languages. We applied rectified linear unit (ReLU) for activation and for the classification we are using softmax. The optimizer is the adam and the loss function is the sparse categorical cross entropy. We passed the data to the model with epochs 20 and batch size 128. The model

gives us percentage by how much that utterance matches to each language. The highest percentage for the given utterance is the final result. Finally, we got the accuracy by the model that can correctly predict the language of given utterance based on the features extracted. We used tensorflow package at the backend.

D. Classification of different langues based on spectrogram and waveform images

We used GPU and tensorflow framework for image classification. Input image size for ResNet-50 is 224×224 and batch size 16 is deployed to process in one iteration. We used 3600 images for training samples, 400 images for validation (close testing) and 400 images for testing samples (open testing) belonging to four classes (Burmese, Chinese, English and Shan). In our experiments, Image-DataGenerator was applied for training pictures with data augmentation, to get more training data by applying different transformations to the existing pictures (rotation, shift, zoom, etc.). Since we had four classes, we chose 'categorical_crossentropy' as the loss function and epochs 50.

We recreate the top layer (classifier) with one additional neuron for the new category and retrain the upper layers with pictures (spectrograms and wave forms). In order to avoid class imbalance, we train the similar number of images.

V. RESULTS AND DISCUSSION

Experimental results of proposed system are explained in this section. In our experiments, language identification using MFCC features was performed on 3200 training data and 800 testing data for Burmese, Shan, English and Chinese languages. We also performed classification on spectrogram images and waveform images with 3200 traing data, 400 validation data and 400 testing data.

Accuracy was calculated for each classification method as in equation 1.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ Number\ of\ predictions} \quad (1)$$

A comparison of the classification accuracies is shown in Table I. Bold numbers indicate the highest scores among the three approaches.

TABLE I
COMPARISON OF IDENTIFICATION ACCURACIES

Method	Accuracy (%)
Identification on MFCC features	**100%**
Identification on wave images	96.5%
Identification on spectrogram images	**100%**

For four languages, the precision, recall and F1 score were presented in order to determine the performance. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Recall is the ratio of correctly predicted positive observations to the all observations in actual class. F1 score is the weighted average of precision and recall. The calculations of precision, recall and F1 score are as in equation 2, 3 and 4.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Postive} \quad (2)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (3)$$

$$F1score = \frac{2 * (Recall * Precision)}{Recall + Precision} \quad (4)$$

TABLE II
PRECISION, RECALL AND F1 SCORE

Features	Precision	Recall	F1 score
MFCC	**1**	1	**1**
waveform images	0.97	0.96	0.96
spectrogram images	1	1	1

As in Table I and II, it can be assumed that the scores of the three approaches are comparable. Identification with MFCC feature and identification using spectrogram achieved the highest accuracy. For Burmese, English, Shan and Chinese languages, these two approaches could correctly classify all audio files of speakers with various accents. For image based identification, classification on the spectrogram images outperfomed over that of waveforms. The language identificaiton on the waveform images could also correctly identify Burmese and Shan voices. However, it contained errors in classification of English and Chinese languages.

VI. CONCLUSION

This work shows the study of three different languge identification models along with their performances in classifying four languages, English, Chinese, Myanmar (Burmese) and Shan (Tai Long). Robust performance can be accomplished with minimal pre-processing. We believe that these models can be extended to classify more languages so long as sufficient, representative training and validation data is available.

REFERENCES

[1] Alexandra Draghici,and Hanna Lukashevich, "A Study on Spoken Language Identification using Deep Neural Networks", 2020-July.
[2] Bartz, C., Herold, T., Yang, H., Meinel, "Language identification using deep convolutional recurrent neural networks" in Neural Information Processing. LNCS, vol. 10639, pp. 880–889. Springer, Cham (2017).
[3] D. Gerhard, Audio signal classification: History and current techniques: Citeseer, 2003.
[4] D. Martinez, O. Plchot, L. Burget, O. Glembek and P. Matejka, "Language recognition in iVectors space", Proc. Interspeech, 2011-Aug.
[5] Fred Richardson, Doug Reynolds and Najim Dehak, "A Unified Deep Neural Network for Speaker and Language Recognition", in INTERSPEECH 2015.
[6] J. J. Burred and A. Lerch, "Hierarchical automatic audio signal classification,"Journal of the Audio Engineering Society, vol. 52, pp. 724-739, 2004.
[7] J. Martinez, H. Perez, E. Escamilla and M. M. Suzuki, "Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques," CONIELECOMP 2012, 22nd International Conference on Electrical Communications and Computers, Cholula, Puebla, 2012, pp. 248-251, doi: 10.1109/CONIELECOMP.2012.6189918. Spectrogram
[8] Julien Boussard, Andrew Deveau, Justin Pyron, "Methods for Spoken Language Identification",in 2017-Dec.
[9] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788-798, May 2011.
[10] Rutuja Ubale, Vikram Ramanarayanan, Yao Qian, Keelan Evanini, Chee Wee Leong, and Chong Min Lee, "Native Language Identification from Raw Waveforms using Deep Convolutational Neural Networks with Attentive Pooling", in IEEE 2019.
[11] Shauna Revay and Matthew Teschke, "Multiclass Language Identification using Deep Learning on Spectral Images of Audio Signals", 2019-May.
[12] T. Giannakopoulos, "PyAudioAnalysis. A Python library for audio feature extraction, classification, segmentation and applications," PloS one, vol. 10, p.e0144610, 2015.
[13] Zazo R, Lozano-Diez A, Gonzalez-Dominguez J, T. Toledano D, Gonzalez-Rodriguez J, "Language Identification in Short Utterances Using Long Short-Term Memory (LSTM) Recurrent Neural Networks", in 2016-Jan.
[14] https://medium.com/analytics-vidhya/how-to-train-your-resnet-the-jindo-dog-50551117381d
[15] https://www.quora.com/What-is-ResNet-50-How-does-it-work-and-how-many-layers-does-it-have
[16] https://www.mathworks.com/help/deeplearning/ref/resnet50.html

lang-identification-revised

10	www.novetta.com Internet Source	1%
11	ceur-ws.org Internet Source	1%
12	"Neural Information Processing", Springer Science and Business Media LLC, 2017 Publication	1%
13	www.ee.iitb.ac.in Internet Source	1%
14	www.waset.org Internet Source	1%
15	www.stargazing.net Internet Source	1%
16	www.eng.biu.ac.il Internet Source	1%
17	kar.kent.ac.uk Internet Source	<1%
18	hdl.handle.net Internet Source	<1%
19	Francesco Mercaldo, Antonella Santone. "Audio signal processing for Android malware detection and family identification", Journal of Computer Virology and Hacking Techniques, 2021 Publication	<1%

20 Soumya Sen, Anjan Dutta, Nilanjan Dey. "Audio Processing and Speech Recognition", Springer Nature, 2019
Publication

<1%

21 Samarpan Guha, Aankit Das, Pawan Kumar Singh, Ali Ahmadian, Norazak Senu, Ram Sarkar. "Hybrid Feature Selection Method based on Harmony Search and Naked Mole-Rat Algorithms for Spoken Language Identification from Audio Signals", IEEE Access, 2020
Publication

<1%

22 "Advanced Concepts for Intelligent Vision Systems", Springer Science and Business Media LLC, 2017
Publication

<1%

23 www.repository.cam.ac.uk
Internet Source

<1%

24 pt.scribd.com
Internet Source

<1%

25 Nikolaos Sideris, Georgios Bardis, Athanasios Voulodimos, Georgios Miaoulis, Djamchid Ghazanfarpour. "Using Random Forests on Real-World City Data for Urban Planning in a Visual Semantic Decision Support System", Sensors, 2019
Publication

<1%

26 "Engineering Applications of Neural Networks", Springer Science and Business Media LLC, 2019
Publication

<1%

27 dokumen.pub
Internet Source

<1%

28 citeseerx.ist.psu.edu
Internet Source

<1%

29 "Advanced Intelligent Systems for Sustainable Development (AI2SD'2019)", Springer Science and Business Media LLC, 2020
Publication

<1%

30 Renxi Chen, Xinhui Li, Shengyang Li. "A Lightweight CNN Model for Refining Moving Vehicle Detection from Satellite Videos", IEEE Access, 2020
Publication

<1%

31 Archi Banerjee, Shankha Sanyal, Souparno Roy, Sayan Nag, Ranjan Sengupta, Dipak Ghosh. "A novel study on perception–cognition scenario in music using deterministic and non-deterministic approach", Physica A: Statistical Mechanics and its Applications, 2021
Publication

<1%