Automatic Language Identification on Audio Signals

28.1.2021
Thursday

Presented by:
Ei Thandar Phyu (ME-IST-2)
Myo Mar Thinn (ME-IST-3)
Khaing Zar Mon (ME-IST-5)
Nang Aeindray Kyaw (ME-IST-6)
May Phyu Khin (ME-IST-7)

Outline

- Motivation
- Introduction
- System Flow
- Methodology
- Data Preparation
- Experimental Setup
- Results and Discussion
- Conclusion

Motivation

- Recently, voice assistants have become a staple in the flagship products of many big technology companies such as Google, Apple, Amazon, and Microsoft.
- To improve user experience on 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.

Introduction

- 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.
- We decided to study language identification of four different languages, namely English, Chinese, Myanmar (Burmese) and Shan (Tai Long).
- Three different approaches were used to carry out the experiments.
  - Language identification from Mel Frequency Cepstral Coefficients (MFCC)
  - Language identification from waveforms of raw audio signals
  - Language identification from spectrograms of raw audio signals

System Flow

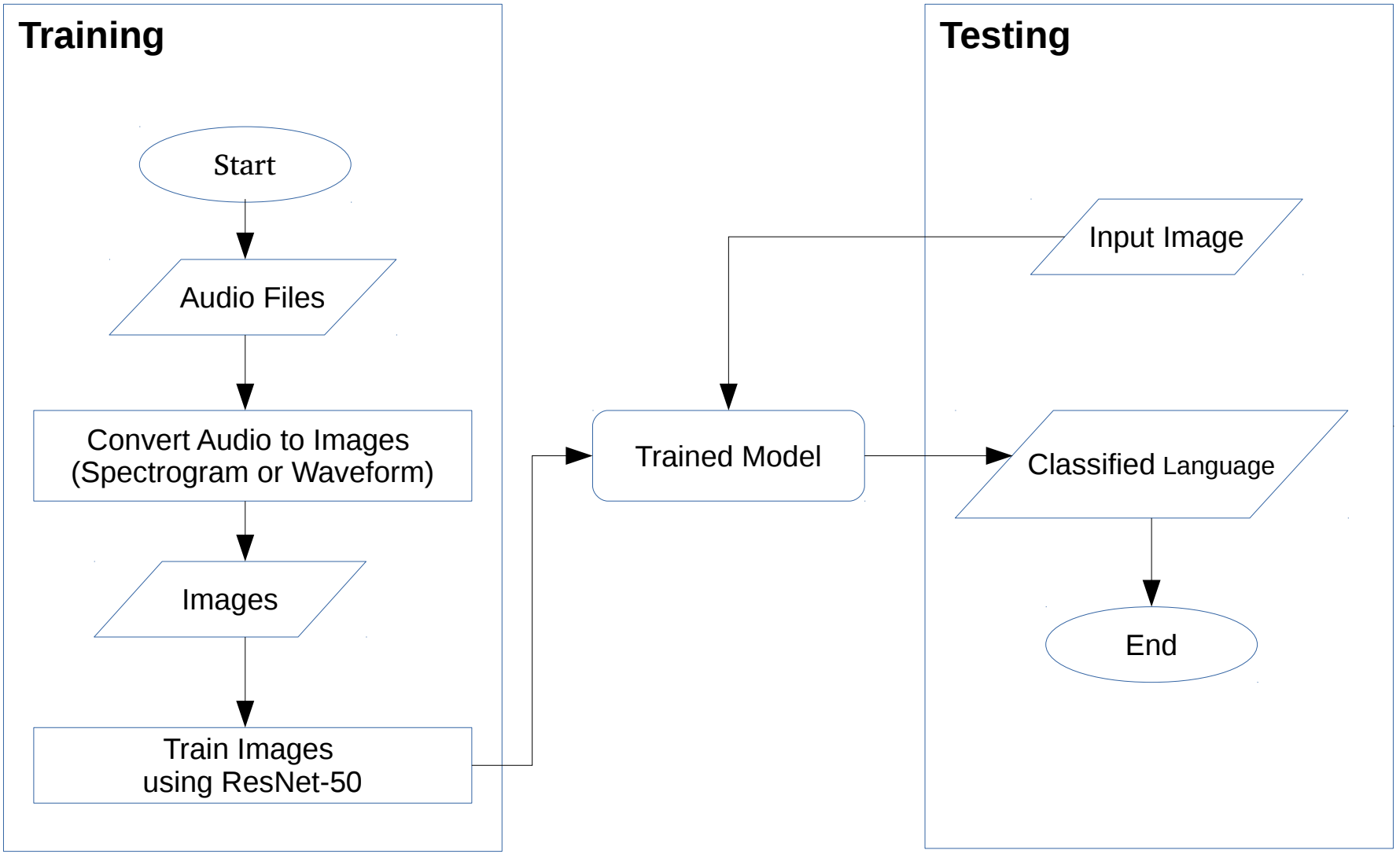

Figure 1. System Flow of Image Based Language Classification

Cont'd

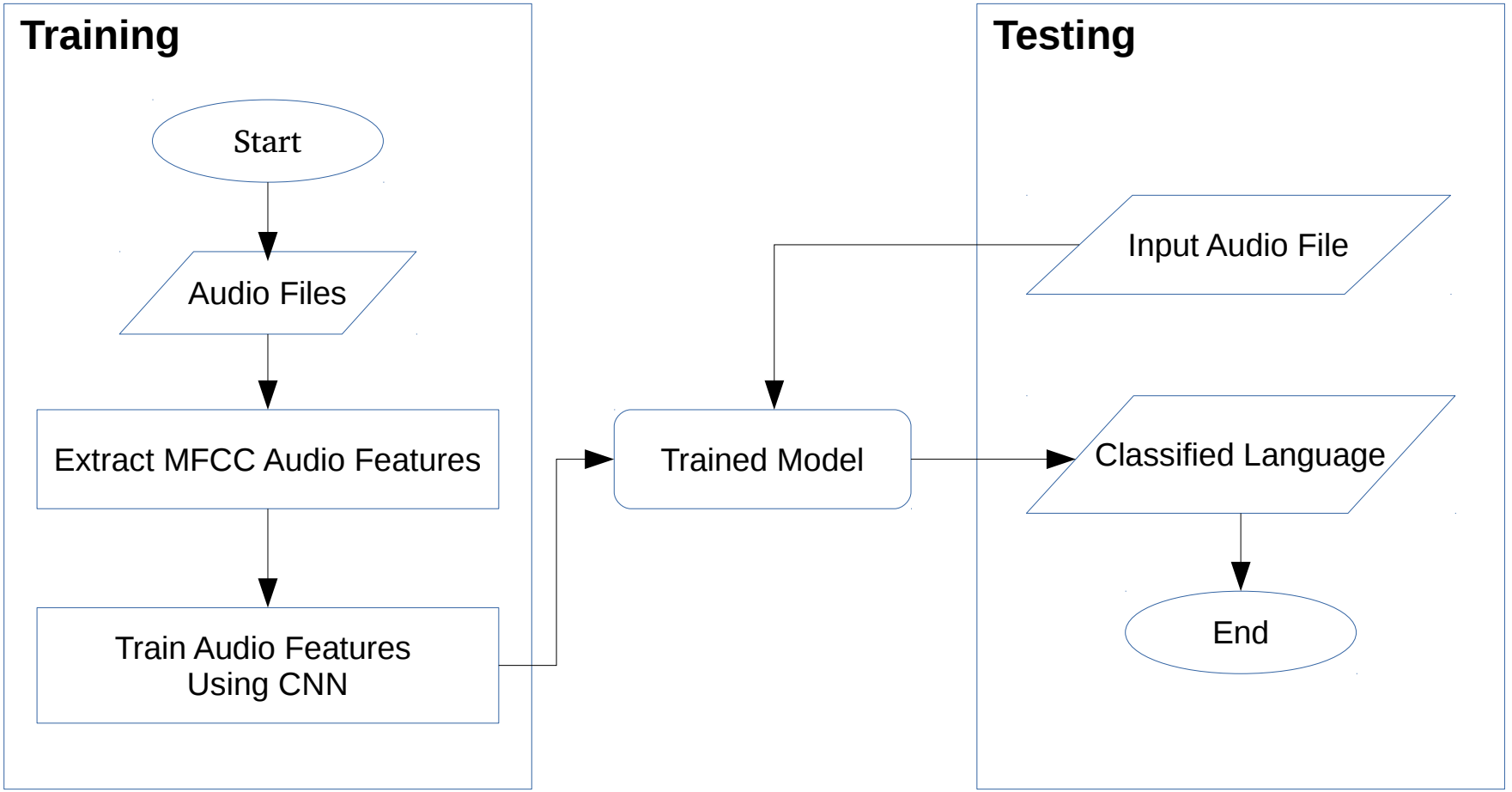

Figure 2. System Flow of Audio Based Language Classification

Methodology

Audio Feature Extaction

- Audio features represent the characteristics of the audio signal and will ultimately decide the class of that signal.
- We used MFCC feature for our experiment.
- 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.

Cont'd

Mel-frequency Cepstral Coefficients (MFCC)

- It is a technique based on the difference of frequencies that the human ear can distinguish.
- MFCCs are the discrete cosine transform coefficients of the mel-scaled log-power spectrum.
- MFCCs have been widely used in speech recognition, speaker clustering and many other audio analysis applications.

Cont'd

Transfer Learning

- A machine learning technique whereby the knowledge gained during training in one problem is used for training in another, similar type of problem.
- 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.
- Transfer learning can aid in training neural networks in considerably less time.

Cont'd

ResNet-50

- ResNet has shown state-of-the-art performance on image recognition tasks.
- The Residual Network with 48 Convolution layers, 1 Max Pool and 1 Average Pool layer.
- The network has learned rich feature representations for a wide range of images.
- ResNet-50 is ideal for speed and accuracy by saving the time and expense of having to do the training from scratch.

Data Preparation

- For Burmese, English, Shan, and Chinese languages, audio clips were gathered by recording male and female speakers with various accents.
- 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.
- There were 1,000 audio files in WAV format for each language respectively.

Cont'd

- All auido files were sampled at a rate of 16kHz with a mono channel.
- Librosa is used for analyzing and extracting audio features of an audio signal.
- For image based classification, the spectrograms of audio files were extracted using the Sound eXchange (SoX, Swiss Army knife of sound processing programs) command line utility.
- The audio files were converted to the waveforms using the FFmpeg cross-platform solution.

Cont’d

Figure 3. Waveform of Chinese Audio File

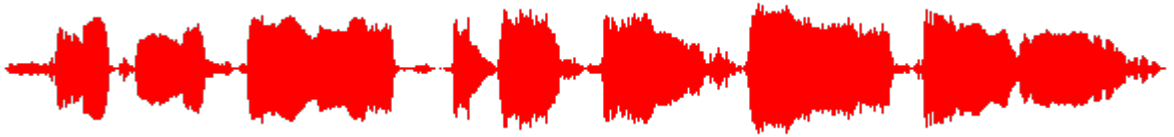

Figure 4. Waveform of English Audio File

Cont'd

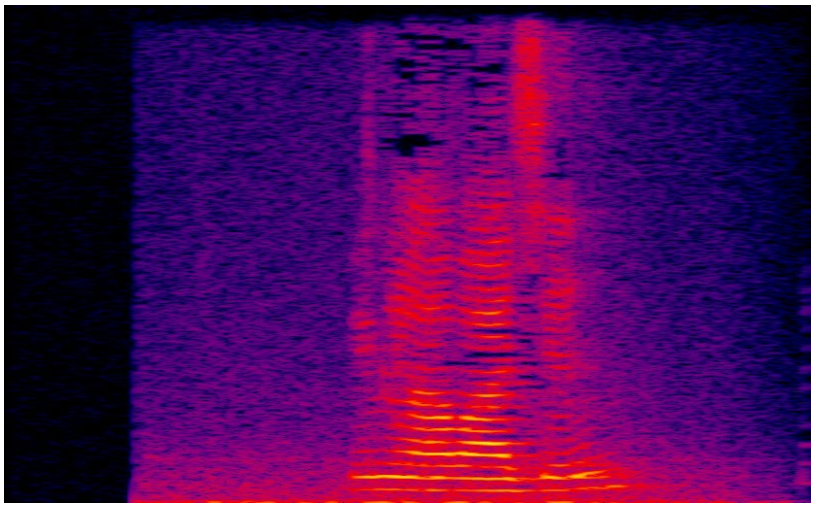

Figure 5. Spectrogram of Burmese Audio File

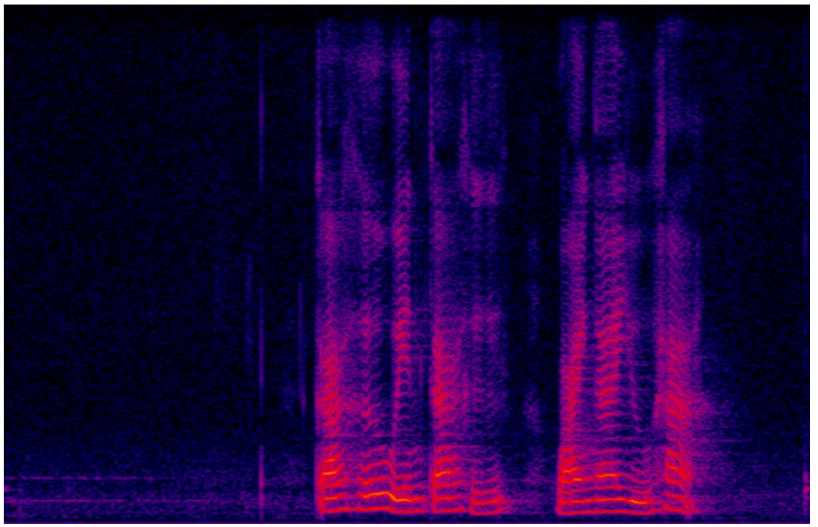

Figure 6. Spectrogram of Shan Audio File

Experimental Setup

Classification based on audio files

- Language identification on MFCC features was performed on 3200 training data and 800 testing data.
- Google Colaboratory (Google Colab) is used as IDE for our experiments.
- The language is predicted based on the Convolutional Neural Network (CNN) with an input layer, an output layer and two hidden layers.

Cont'd

- We applied rectified linear unit (ReLU) for activation and used softmax for classification.
- The optimizer is the adam and the loss function is the sparse categorical cross entropy.
- The model is trained with epochs 20 and batch size 128.
- We used tensorflow package at the backend.

Cont'd

Classification based on spectrogram and waveform images

- We used 3,600 training images, 400 validation images (close test) and 400 testing images (open test) for each classification on spectrogram images and waveform images.
- The models are trained on GPU using tensorflow framework and Residual Network (ResNet-50).
- The classifier with one additional neuron for the new category is recreated.

Cont’d

- The upper layers are retrained with spectrogram and waveform images.
- The input image size of ResNet-50 is 224-by-224 and batch size is 16.
- ImageDataGenerator was applied for training pictures with data augmentation, to get more training data by applying different transformations to the existing pictures.
- 'categorical_crossentropy' is applied as the loss function and each experiment is trained with epochs 50.

Results and Discussion

- For audio classification of four languages, the precision, recall and F1 score were presented in order to determine the performance.
- According to the experimental results, the identification using MFCC features and the identification using spectrograms could correctly classify all Burmese, Chinese, English and Shan languages.

Features	Precision	Recall	F1 Score
MFCC	**1**	**1**	**1**
Waveforms	**0.97**	**0.96**	**0.96**
Spectrograms	**1**	**1**	**1**

Table 1. Precision, Recall and F1 Score of three classification models

Conclusion

- The study of three different languge identification models is shown 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.
- These models can be extended to classify more languages so long as sufficient, representative training and validation data is available.

References

- Alexandra Draghici,and Hanna Lukashevich, "A Study on Spoken Language Identification using Deep Neural Networks", 2020-July.
- 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.
- D. Gerhard, Audio signal classification: History and current techniques: Citeseer, 2003.
- D. Martinez, O. Plchot, L. Burget, O. Glembek and P. Matejka, "Language recognition in iVectors space", Proc. Interspeech, 2011-Aug.
- Fred Richardson, Doug Reynolds and Najim Dehak, "A Unified Deep Neural Network for Speaker and Language Recognition", in INTERSPEECH 2015.

Cont'd

- J. J. Burred and A. Lerch, "Hierarchical automatic audio signal classification,"Journal of the Audio Engineering Society, vol. 52, pp. 724-739, 2004.
- Julien Boussard, Andrew Deveau, Justin Pyron, "Methods for Spoken Language Identification",in 2017-Dec.
- 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.
- 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.

Cont’d

- Shauna Revay and Matthew Teschke, "Multiclass Language Identification using Deep Learning on Spectral Images of Audio Signals", 2019-May.
- T. Giannakopoulos, "PyAudioAnalysis. A Python library for audio feature extraction, classification, segmentation and applications," PloS one, vol. 10, p.e0144610, 2015.
- 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.

Thank You