Language Identification Challenge for TEDx Talks

April 25, 2025

Language Identification (LID) systems from voice are classification models that predict the spoken language from a given audio recording. The LID systems can facilitate the process of any speech processing system such as speech recognition (ASR) or speech translation systems. In speech-based assistant systems, LID works as a first step by selecting the appropriate grammar from a list of available languages for further semantic analysis. Also, these models can be employed in call centers in order to redirect an international user to an operator who is fluent in that identified language.

1 Objective

The objective of this project is to use machine learning methods for constructing a LID model which can discriminate 4 languages; English, French, Arabic, Japanese. There are 2 expected phases. The first phase is constructing a classifier. It is expected to compare the performance of different models, optimize the hyperparameters, and practice of finding the best model. The second phase is the evaluation of the model in a simulated situation of real life deployment. The objective is to understand the challenge of generalization. It's also expected to analyze the result of the model's performance and make hypothesis about the weak and strong aspect of models. The competition among models' accuracy can provide a better understanding of performance.

The main task of this small project is to create a classification model which generates a label as the spoken language for a given audio file in *.wav* format. It is expected to train different models with different hyperparameters and find the best model. Afterward, predicting the label of samples in Test set.

Bonus: Use ensemble learning; train several models which provide several predicted labels per file, aggregate the labels (majority votes).

2 Dataset

The provided dataset has been collected from TEDx talks YouTube for the Language Identification task from audio. The samples are recorded audio of speaker speech from available TEDx talks videos. In order to have a standard sample's type, they follow below convention.

- The length of recorded audio files should be around 5 seconds (5.00 - 5.99 seconds).
- The format of audio files should be **.wav*.
- The sample rate of recording files should be 16 kHz (in mono format).

2.1 Training/Development set

A repository contains recording files (587 files) in the standard format (*.wav, 16kHz, mono, 5-6 seconds) and a *.txt file with 4 information (separated by ,) for each recorded file (one file per line) has been provided.

- The 1st column is the name of *.wav file
- The 2nd column is the URL address of YouTube video
- The 3rd column is the starting time of recording from YouTube video

- The 4th column is the label (language) of recorded speech (EN, FR, AR, JP)

Link to Training/Development set

2.2 Test set

A repository contains recording files (832 files) in the standard format (*.wav, 16kHz, mono, 5-6 seconds) and a *.csv file with 2 columns for the file names and the corresponding predicted labels by your classification model. Link to Test set. The first line of the csv file must be "ID, Label" as the column's name.

3 Evaluation of practical course (TP)

3.1 Kaggle

Register (sign up) an account on *kaggle.com* use below link to access to the ENSIM LID Challenge. Link to ENSIM LID challenge 2025

Generate a *txt* file with 2 columns for the file names (ID) and the corresponding predicted labels (Label) by your best model for the Test set. Submit this model on the *kaggle*. An example of predicted label file (txt file) for Test set is available on the UMTICE. It is expected that your model performance would be significantly higher than random classifier (25%).

In the Describe submission section, enter the following information in the same pattern (separated by ,) "Your NAME, Your TP group, Your Model name, version".

Students with the best models get extra points.

3.2 Presentation

Prepare a presentation slides in PDF format (MAX 10 slides) with important points and upload on the UMTICE Your name and your TP group name on the first slide.

3.3 The evaluation chart

All below parts should be reported by right performance values with a

- Partitioning of data into train/development or using k-fold cross validation : 10
- Training a GNB : 5
- Training SVM models with different hyperparameters : 10
- Training Random Forest models with different hyperparameters : 10
- Training MLP models with different hyperparameters : 10
- Comparing different feature extractors (at least 2) : 5
- Comparing the results with a dummy model : 5
- Prediction of test labels and submission to the Kaggle : 5
- Comparing the result of Kaggle and the expected performance on the development set partition : 5
- Comparing the impact of training size : 5
- Calculating the confidence interval to find the significance of performances difference : 10
- Reporting the results in table or figure instead of screenshot or text, and note a conclusion for each figure : 5
- Analysing the best model performance with different metrics (like weighted/unweighted accuracy or f1-score) on different classes (confusion matrix) : 5

- Presentation in the last session (ability to focus on the important parts of work in a time) : 10
- Extra points:
  The impact of PCA on feature extraction : 5
  Using ensemble learning with difference classifiers : 10
  Using another classifier that has not been asked : 5
  Using data augmentation and its impact : 10
  Perspectives and proposition of future plan to improve the accuracy (justify with evidence) : 5

4 More information

- Feature extraction using librosa
- A project report for music classification
- An example code for classification of animal sound