# YZV303(2)E/BLG561E Deep Learning Project Proposal Project Title*

Nurullah Eren Acar
*Artificial Intelligence and Data Engineering*
*Istanbul Technical University*
acarn22@itu.edu.tr
150220310

Ömer Faruk Satık
*Artificial Intelligence and Data Engineering*
*Istanbul Technical University*
satik21@itu.edu.tr
150210330

## I. Project Description

This project aims to develop a language detection system from speech using deep learning techniques. In academic literature, this topic is often referred to as Spoken Language Identification(SLI or LID). SLI can be viewed as a multiclass classification problem of assigning natural language labels onto audio signals of arbitrary length, based on the unique acoustic structure of each language (Muthusamy and Cole 1992) [1]. The focus of the project is on accurately identifying the spoken language from audio recordings scrapped from web. By leveraging spectrogram-based data representations and neural network architectures, this project will explore the phonetic and acoustic patterns of different languages. The system will focus on six languages: English, Russian, Turkish, Chinese, Arabic and Hindi. The project aims to contribute in speech processing field where multilingual environments are existed.

## II. Problem Definition

LID technology is needed in many applications such as multilingual conversational systems, spoken language translation, multilingual speech recognition, and spoken document retrieval [2]. It is also a topic of great importance in the areas of intelligence and security, where the language identities of recorded messages and archived materials need to be established before any information can be extracted from them [2]. For these and many similar reasons, humanity needs LID systems in many areas of life. However, language identification from speech task can be quite challenging because of the relations to the phonetic, acoustic and grammatical structure of the language. It can be difficult for both the humans and machines to perceive the distinctive features between different languages. Correctly modeling the sound can reveal patterns that couldn't be perceived even by the human ear, helping in creating a reliable and automated language detection system. For these reasons and purposes, using artificial intelligence solutions to distinguish languages from each other can make a lot of sense.

## III. Dataset

For this project, creating a well planned dataset is really important, being the backbone of our training process. We will focus on a wide range of speakers among the six languages: English, Russian, Turkish, Chinese, Arabic, and Hindi. YouTube provides a tremendous amount of data in case of audio. Interviews, podcasts, lectures type of videos will be used to gather data due to their rich linguistic variety and good audio quality.

We will develop a sophisticated and advanced scraping technique which will systematically extract these audio recordings. Noise free audio will be prioritized with clear speech content. Manual checks will also be done to ensure that the audio recordings meet the needed quality standarts.

After the audio recordings are collected they will undergo a preprocessing pipeline. All audio files will be converted into a specified uniform file format and sample rate in order to maintain uniformity in representations. Longer audio recordings will be converted into small chunks to pave the way for an easier training process. Each audio segment will be converted into a spectogram using Fast Fourier Transform, giving us a two dimensional representation for the frequency-time domain. Like in the image processing, data augmentation techniques such as time strecthing, pitch shifting and noise injection will be used to improve dataset diversity. Also Variational Autoencoder approach might offer a good solution for augmenting the data.

## IV. Methodology

Methods we are going to use for our project focuses on using state of the art deep learning techniques to create a reliable language detection system. First we create a dataset using the semi-automated system in order to collect data across the content avaiable on YouTube among the wide range of languages we are planning to do our project on. After that we plan to use Voice Activity Detection (VAD) models to isolate voice segments which include people speaking to later feed them into our data preprocessing pipeline where the audio gets segmented into smaller chunks and spectrograms.

For the modeling part, we plan to experiment across multiple neural network architectures. Convolutional Neural Networks(CNNs) has been used widely in various application areas, including image classification and speech recognition [3]. In these domains, A CNN has achieved state-of-the-art levels of performance [3]. Thus, CNNs are a good candidate for extracting spatial features from spectograms, making use of their ability to effectively process image like data.

To capture the unique characteristics and nature of speech data, Recurrent Neural Networks (RNNs) may also be employed helping the model to learn the different patterns that occur in across the languages. Over the last few years, RNNs and in particular RNNs based on LSTM have been successfully applied to a wide range of classification tasks for which the discriminative information is embedded in a sequence [4]. Also we will explore transformer based models which feature attention mechanisms to make use of the complex relationships between frequency and time domains. These models have shown that they have superior performance in speech and audio processing tasks recently.

Final system decision will depend on model performances that will be evaluated during training and testing. Hybrid systems like Recurrent Convolutional Neural Networks may be implemented which combine the strenghts of different architectures making use of the both worlds. The entire system will focus on achieveing the maximum accuracy by leveraging the advantages of quality dataset creation, data augmentation, hyperparameter tuning, tremendeous amounts of evaluation. Additionally, precision and recall metrics will be important criteria in evaluating model performance.

## REFERENCES


[1] Lindgren, M. (2020). Deep learning for spoken language identification (Master's thesis).

[2] Lee, CH. (2008). Principles of Spoken Language Recognition. In: Benesty, J., Sondhi, M.M., Huang, Y.A. (eds) Springer Handbook of Speech Processing. Springer Handbooks. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-49127-9_39

[3] Alashban, A. A., Qamhan, M. A., Meftah, A. H., & Alotaibi, Y. A. (2022). Spoken Language Identification System Using Convolutional Recurrent Neural Network. Applied Sciences, 12(18), 9181. https://doi.org/10.3390/app12189181

[4] Gelly, G., & Gauvain, J. L. (2017, August). Spoken Language Identification Using LSTM-Based Angular Proximity. In Interspeech (pp. 2566-2570).