**Chapter 1 : Introduction**

The problem of automatic language identification (LID) can be defined as the process of automatically identifying the language of a given spoken utterance. The first step in a language processing pipeline is always LID. LID is used either as a standalone task or as a preprocessing step, capturing the first seconds of the recording and processing it in order to transfer the control to the appropriate next stage; e.g. speech recognition systems, multilingual translation systems or call-centers (e.g., emergency calls) routing, where the response time of a native operator might be critical. LID is also a topic of great importance in 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.

Our project aims at spoken language identification in speech audio samples using Deep Learning Models (DNNs). Speech audio recognition features well-defined, clear voices and shows very little background noise. This is in contrast to song audio which combines background music, instruments and the singer’s voice into a complex mix. Even though several high level approaches based on phonotactic and prosody are used as meaningful complementary sources of information, nowadays, many state-of-the-art LID systems still include or rely on acoustic modelling. While previous works on neural networks applied to LID report results using shallow architectures or convolutional neural networks, in this study, we propose the use of DNNs as a new method to perform LID at the acoustic level. Recent breakthroughs in signal processing and acoustic modelling have shown the promise of deep learning models.

Motivated by those results and also by the discriminative nature of DNNs, we adapt DNNs to work at the acoustic frame level to perform LID. Particularly, in this work, we extract the acoustic features from an audio signal, mfcc and their first and second order differentials to be exact which has been divided into frames and compute their mean and variance across several frames in order to account for the fluctuation in the speech signal. This results in stacking of average features across several consecutive frames known as context windows which also result in much superior performance. The mel-frequency cepstrum is a representation of an audio signal on the mel scale, a nonlinear mapping of frequencies that down-samples higher frequencies to imitate the human ear’s ability to process sound. Delta features and the Delta Delta Features are the first and second time derivatives of the cepstral coefficients, capturing the change of the cepstral features over time. Feature Selection using Chi Square Statistics is also applied to obtain the best possible results as well as reduce dimensionality of the input. We compare the obtained results with a baseline Support Vector Machine (SVM) based system trained from exactly the same acoustic features.

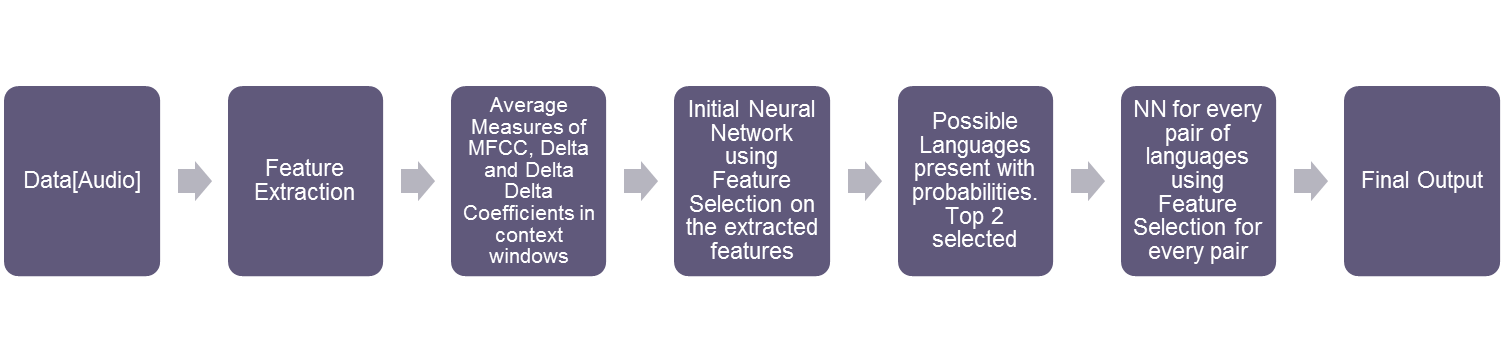
The remaining thesis is organized as follows. In chapter 2, previous work done and the current state of the art in this area are detailed, along with the limitations and the roadblocks encountered in this field. Chapter 3 provides and overview of the technologies used and the theoretical concepts behind using the frameworks or tools. Chapter 4 discusses about our contribution to this field of research. A detailed experimental framework and the course of action is elaborated. Chapter 5 provides an experimental evaluation of our proposed approach, and the findings of the experiment are reported and analyzed. Chapter 6 summarizes the thesis with conclusions and avenues of future work. 

Fig **{number}**Workflow of the project