Karlsruhe Institute of Technology

Online Neural Network-based Language Identification

Master's Thesis of

Daniel H. Draper

at the Department of Informatics
Institute for Anthropomatics and Robotics

Reviewer:	Dr.-Ing. Sebastian Stüker
Second reviewer:	
Advisor:	M.Sc. Markus Müller

12. December 2016 – 11. May 2017

Karlsruher Institut für Technologie
Fakultät für Informatik
Postfach 6980
76128 Karlsruhe

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

Karlsruhe, 12th of May, 2017

. . . .

(Daniel H. Draper)

Abstract

Zusammenfassung

Contents

List of Figures

List of Tables

1 Introduction

2 Preliminary Definitions

In the following chapter we want to define and explain terms and concepts used throughout this thesis as well as give an outlook to related work and the general language identification approaches.

2.1 Related Work

In this section we take a look at related work that shows different approaches of identifying Language in spoken speech and describe the differences between their work and our approach.

3 Language Identification Tasks

This chapter introduces the datasets used to train the networks employed in this approach. While Language Identification is applicable in many different scenarios, in this thesis the focus lies on trying to establish a low-latency online approach for recognizing the spoken language in a university-lecture environment. Because finding a suitable test setup for online data retrieval is hard the data used was cut to short lengths to make an evaluation as to correctness of the recognition possible in an ”online-like” scenario.

This means that the output of the net is evaluated after short samples of speech and therefore can be seen as indicative of online performance of the neural net.

3.0.1 Euronews 2014

Our first data set we retrieved from Euronews [1] 2014. Euronews is a TV channel that is broadcast in 13 different languages simultaneously both on TV and over the Web. The first data corpus includes our 10 language (Arabian, German, Spanish, French, Italian, Polish, Portugese, Russian, Turkish and English) with about 20 hours of data per language provided overall. Details of this breakdown can be seen in table 3.0.1.

Language	Number of Speakers	Length overall
Arabian	1055	
German	928	
Spanish	932	
French	1016	
Italian	935	
Polish	1229	
Portugese	1062	
Russian	958	
Turkish	957	
English	928	
Overall	10000	

Table 3.1: The Euronews corpus speaker breakdown with total utterances length

The speaker list was then split into three smaller datasets: the train set, development set and test set using the common Simple Random Sampling. The sizes were 80% train set, and 10% for both the development and test set. Table **??** shows the split data for the three sets.

[1] Euronews: http://www.euronews.com/

3.0.2 Lecture Data

3.0.3 European Parliament

4 Feature Preprocessing

This chapter deals with the feature preprocessing used to form normal speech into feature vectors to be understood by neural networks. The setup we used is based on the standard capabilities of the Janus Recognition Toolkit[1]. It is then run through a six layer Automatic Speech Recognition network that was pretrained on 10 languages. The 2nd last layer of the ASR net is a Bottleneck feature layer, where the feature vectors are extracted and then used as input for the trained Language Identification Network. The following sections describe this first network in detail and explain the Feature Preprocessing.

4.0.1 Feature Access

4.0.2 Feature Description

The extracted ADC features from the audio files are then used for further preprocessing. We first use a standard mel filter bank to extract only the necessary coefficients from the ADC0 feature.

First a spectrum is applied to the ADC0 Feature, therefore calculating the Fast Fourier Transformation of the Digitalized Signal.

[1] Janus Recognition Toolkit(JRTk): http://isl.anthropomatik.kit.edu/cmu-kit/english/1406.php

5 Conclusion