5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

Investigating Bilingual Deep Neural Networks for Automatic Recognition of Code-switching Frisian Speech

Emre Yılmaz*, Henk van den Heuvel, David van Leeuwen

CLS/CLST, Radboud University, Nijmegen, Netherlands

Abstract

In this paper, a code-switching automatic speech recognition (ASR) system built for the Frisian language is described. Frisian is mostly spoken in the province Fryslân which is located in the north of the Netherlands. The native speakers of Frisian are mostly bilingual and often code-switch in daily conversations due to the extensive influence of the Dutch language. In the scope of the FAME! Project, the influence of this unforeseen language switching on modern ASR systems will be investigated with the objective of building a robust recognizer that can handle this phenomenon. For this purpose, in this work, we design a bilingual deep neural network (DNN)-based ASR system and investigate the impact of bilingual DNN training in the context of code-switching speech.
© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license
(http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords: Automatic speech recognition, bilingual DNN, code-switching, low-resourced languages, Frisian

1. Introduction

Language interaction in minority languages spoken in multilingual countries has been researched in the field of linguistics for more than 30 years[1,2,3]. This interaction occurs in the form of phonological, morphological, syntactic and lexical changes consequent to various linguistic phenomena such as word borrowing, interference and relexification. One prominent mechanism induced in the interacting languages is code-switching (CS) which is defined as the continuous alteration between two languages in a single conversation.

CS is mostly noticeable in some minority languages influenced by the majority language or majority languages that have been influenced by globally influential languages such as English and French. Despite the well-established research line in linguistics, robustness of speech-to-text systems against CS and other kinds of language switches have recently received some interest resulting in some robust acoustic modeling[4,5,6,7,8] and language modeling[9,10,11] approaches for CS speech.

---

* Corresponding author. Tel.: +31243612055
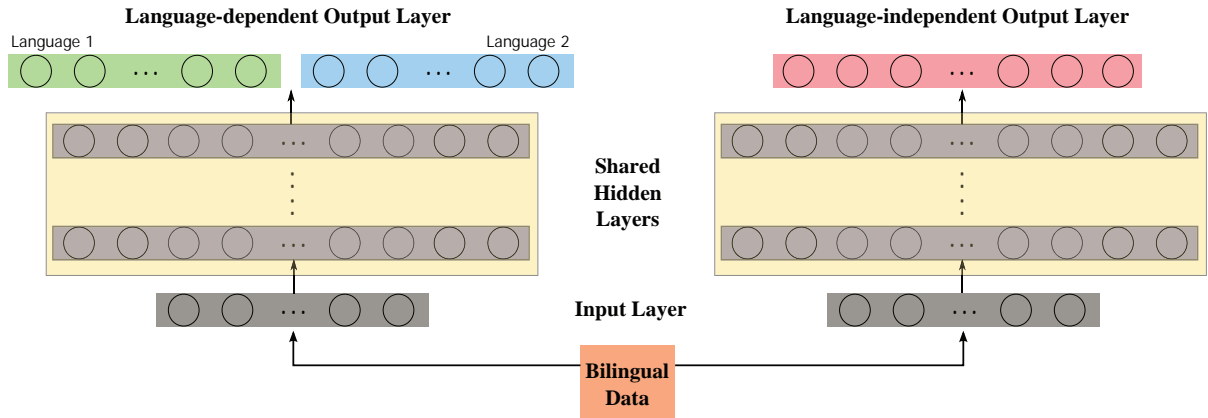 *E-mail address:* e.yilmaz@let.ru.nl

Fig. 1. Language-dependent and language-independent bilingual DNN architecture

Language identification and/or diarization is a relevant task for the automatic speech recognition (ASR) of CS speech[12,13,14,15]. One fundamental approach is to label speech frames with the spoken language and perform recognition of each language separately using a monolingual ASR system at the back-end. These systems have the tendency to suffer from error-propagation between the language identification front-end and ASR back-end, since language identification is still a challenging problem especially in case of intra-sentence CS. To alleviate this problem, single-pass ASR approaches, which do not directly incorporate a language identification system, have also been proposed yielding promising results[5,8].

Multilingual training of deep neural network-based ASR systems has provided some improvements in the recognition accuracies for both low- and high-resourced languages[16,17,18,19,20,21,22,23,24]. Some of these techniques incorporate multilingual DNNs for feature extraction[25,16,26,21]. Training DNN-based acoustic models on multilingual data to obtain more reliable posteriors for the target language has also been investigated[19,20,24]. ASR systems using DNN-HMM models can employ language-independent or language-dependent phonetic alphabets depending on the purpose of the multilingual training.

In this work, we describe a bilingual deep neural network (DNN)- based ASR system which is designed to recognize both Frisian and Dutch. By investigating different bilingual DNN architecture, we aim to get more insight into developing a more versatile acoustic modeling scheme coping with the language switches. Specifically, both phone-dependent and phone-independent bilingual DNN training approaches are applied on the novel Frisian database and the recognition performance of both systems is compared in order to have a better understanding how far a CS ASR system can benefit from phone merging and cross-language knowledge transfer by using shared hidden layers. To the best of our knowledge, the impact of bilingual DNN training on CS speech data has never been explored. To be able to make a fair comparison of both architecture, the proposed ASR system uses a bilingual lexicon and a bilingual language model trained on a text corpus containing Frisian, Dutch and mixed-language sentences.

This paper is organized as follows. Section 2 summarizes the novel Frisian database that has recently been collected for CS and longitudinal speech research. Section 3 summarizes the DNN-based ASR and Section 4 describes the CS ASR system. The experimental setup is described in Section 5 and the recognition results are presented in Section 6. Section 7 concludes the paper.

2. FAME!: Frisian Radio Broadcast Database

The FAME! speech database has been collected in the scope of the Frisian Audio Mining Enterprise Project. This project aims to build a spoken document retrieval system for the disclosure of the archives of Omrop Fryslân (Frisian Broadcast) covering a large time span from 1950s to the present and a wide variety of topics. Omrop Fryslân is the regional public broadcaster of the province Fryslân with a radio broadcast archive containing more than 2600 hours of recordings. The FAME! speech database contains a small subset of these radio broadcasts and it is the first spoken Frisian database of high recording and annotation quality. The recordings selected for the database