Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 117 (2017) 81–88

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5–6 November 2017, Dubai, United Arab Emirates

Development of the Arabic Loria Automatic Speech Recognition system (ALASR) and its evaluation for Algerian dialect

Mohamed Amine Menacer[a], Odile Mella[a], Dominique Fohr[a], Denis Jouvet[a], David Langlois[a], Kamel Smaïli[a]

[a]*Loria, Campus Scientifique, BP 239, Vandoeuvre Lès-Nancy, 54506, France*

Abstract

This paper addresses the development of an Automatic Speech Recognition system for Modern Standard Arabic (MSA) and its extension to Algerian dialect. Algerian dialect is very different from Arabic dialects of the Middle-East, since it is highly influenced by the French language.

In this article, we start by presenting the new automatic speech recognition named ALASR (Arabic Loria Automatic Speech Recognition) system. The acoustic model of ALASR is based on a DNN approach and the language model is a classical n-gram. Several options are investigated in this paper to find the best combination of models and parameters. ALASR achieves good results for MSA in terms of WER (14.02%), but it completely collapses on an Algerian dialect data set of 70 minutes (a WER of 89%). In order to take into account the impact of the French language, on the Algerian dialect, we combine in ALASR two acoustic models, the original one (MSA) and a French one trained on ESTER corpus. This solution has been adopted because no transcribed speech data for Algerian dialect are available. This combination leads to a substantial absolute reduction of the word error of 24%.

Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.

Keywords: speech recognition; Algerian dialect; language model; acoustic model; MSA.

1. Introduction

Modern Standard Arabic, generally referred as MSA (Alfus'ha in Arabic), is the variety of Arabic that was retained as the official language in all Arab countries, and as a common language between them. Standard Arabic is not acquired as a mother tongue, but rather it is learned as a second language at school and through exposure to formal broadcast programs (such as the daily news), religious practice and newspapers reading. MSA is characterized by a complex morphology and a rich vocabulary, since it is a very old language, and it can still be understood in its original form in a better way than French or English. It is an inflectional and agglutinative language; let's recall that compared to English, an Arabic word (or more rigorously a lexical entry) can sometimes correspond to a whole English sentence.

* Corresponding author. Tel.: +33-755-260-590.
E-mail address: mohamed-amine.menacer@loria.fr

1877-0509
Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.
10.1016/j.procs.2017.10.096

The grammatical system of Arabic language is based on a root and pattern structure, which uses more than 10000 roots and 900 patterns [5]. Arabic words are derived from roots by using patterns.
However, most of Arab people do not use MSA in their daily conversations, since their mother tongue is an Arabic dialect, which is derived from MSA. Furthermore, for each region, there is one or several dialects influenced by the history of the region itself. New words borrowed from English, Turkish, Spanish, Italian or French are integrated in the vocabulary of these dialects. Some of these new words are used such as in the original language, but others are altered in order to respect the morphological structure of the Arabic language. For instance, the French word *casseroles* (*pans*), which is pronounced /k a s r o l/, has been borrowed and integrated in Algerian dialect. It is written in this dialect as follows: كسرونَات, which is uttered /k a s r uː n aː t/.

Our objective in this paper is to present the development of an Automatic Speech Recognition (ASR) system for MSA (namely ALASR: Arabic Loria Automatic Speech Recognition), to evaluate its performance on MSA data and then to test it on a corpus that is not entirely MSA. This corpus is composed of Algerian dialect sentences extracted from PADIC [17].
The development of such system requires to take into account the characteristics of MSA (highly inflectional and agglutinative language), several studies have dealt with some of these issues. In [10], [11], [19] and [9], the authors proposed to use morphological segmentation to reduce the size of the lexicon. In some of these works they maintain the quality of the ASR and for others they improve the quality of their systems.

The absence of diacritics in Arabic texts is a serious issue for many applications in Natural Language Processing (NLP). For every Arabic root which is not vocalized, the ASR system has to consider all the possibilities of pronunciations or has to restore the diacritics. For example, the word قبل can be pronounced such as: قَبلَ /q a b l a/ (*before*), قَبِلَ /q a b i l a/ (*accepted*) or قُبِّلَ /q u bː i l a/ (*was kissed*) and it also has other diacritizations, which are not listed here.
It is shown in [10] that using short vowels in the training transcripts and applying a morphological decomposition on the vocabulary, improved the Word Error Rate (WER).
In speech recognition, the bulk of works proposed in the literature is intended for MSA, while only few works are dedicated to Arabic dialects. This is mainly due to the lack of Arabic dialect data [25, 6]. The efforts made, until now, to develop ASR for Arabic dialects concern those considered as close to MSA, namely Iraqi [3], Egyptian [7], Qatari [13] and Leventine [23, 12, 21]. However, there are few ASR systems for Maghrebi dialects especially those used in Algeria. The main characteristic of Algerian dialect is the code-switching, meaning a speaker alternates between MSA and the native dialect, which is influenced by French, Berber and Turkish.
In this work, we decided to test ALASR system on Algerian dialect since it is difficult to develop an entire ASR due to the lack of data necessary for training the acoustic model. To deal with this issue, some changes on the acoustic and language models have been made, in order to make ALASR system doing its best on Algerian dialect.
In the next section, a description of ALASR system is presented. Section 3 gives an overview of the differences between MSA and Algerian dialect and presents the modifications adopted in order to make ALASR usable with Algerian dialect. Experimental results are discussed in Section 3.4.

2. ALASR: Arabic Loria Automatic Speech Recognition system

A speech recognition system needs at least two components: an acoustic model and a language model. In the following, each component is described by presenting the different steps required to train each model. Training necessitates two kinds of data: acoustic and textual, which are presented in the following.

2.1. Acoustic model

The development of the acoustic model is based on Kaldi [20], which is a state-of-the-art toolkit for speech recognition based on Weighted Finite State Transducers [18]. The recipe used relies on 13-dimensional Mel-Frequency Cesptral Coefficients (MFCC) features with their first and second order temporal derivatives, which leads to 39-dimensional acoustic features. For Arabic, 35 acoustic models (28 consonants, 6 vowels and silence) are trained. The emission probabilities of the HMM models are estimated by DNN (namely DNN-HMM). The DNN-HMM are trained