Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 128 (2018) 32–37

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Natural Language and Speech Processing, ICNLSP 2015

Automatic Speech Recognition Errors Detection and Correction: A Review

Rahhal Errattahi[a,*], Asmaa El Hannani[a], Hassan Ouahmane[a]

[a]*Laboratory of Information Technologies, National School of Applied Sciences, University of Chouaib Doukkali, El Jadida - Morocco*

Abstract

Even though Automatic Speech Recognition (ASR) has matured to the point of commercial applications, high error rate in some speech recognition domains remain as one of the main impediment factors to the wide adoption of speech technology, and especially for continuous large vocabulary speech recognition applications. The persistent presence of ASR errors have intensified the need to find alternative techniques to automatically detect and correct such errors. The correction of the transcription errors is very crucial not only to improve the speech recognition accuracy, but also to avoid the propagation of the errors to the subsequent language processing modules such as machine translation. In this paper, basic principles of ASR evaluation are first summarized, and then the state of the current ASR errors detection and correction research is reviewed. We focus on emerging techniques using word error rate metric.

Keywords: Automatic Speech Recognition; ASR Error Detection; ASR Error Correction; ASR evaluation;

1. Introduction

Automatic Speech Recognition (ASR) systems aims at converting a speech signal into a sequence of words either for text-based communication purposes or for device controlling. The purpose of evaluating ASR systems is to simulate human judgement of the performance of the systems in order to measure their usefulness and assess the remaining difficulties and especially when comparing systems. The standard metric of ASR evaluation is the Word Error Rate, which is defined as the proportion of word errors to words processed.

ASR has matured to the point of commercial applications by providing transcription with an acceptable level of performance which allows integration into many applications. In general, ASR systems are effective when the conditions are well controlled. Nevertheless, they are too dependent on the task being performed and the results are far from ideal, and especially for Large Vocabulary Continuous Speech Recognition (LVCSR) applications. This later still one of the most challenging tasks in the field, due to a number of factors, including poor articulation, variable

* Corresponding author. Tel.: +212-523-344-822 ; fax: +212-523-394-915.
E-mail address: errattahi.r@ucd.ac.ma

10.1016/j.procs.2018.03.005

speaking rate and high degree of acoustic variability caused by noise, side-speech, accents, sloppy pronunciation, hesitation, repetition, interruptions and channel mismatch, and/or distortions. To deal with all these problems, there has been a plethora of algorithms and technologies proposed by the scientific communities for all steps of LVCSR over the last decade: pre-processing, feature extraction, acoustic modeling, language modeling, decoding and result post-processing. Nevertheless LVCSR systems are not yet robust with error rates of up to 50% under certain conditions [21],[8].

The persistent presence of ASR errors motivates the attempt to find alternative techniques to assist users in correcting the transcription errors or to totally automate the correction process. Manual errors correction is often tedious and time consuming. Hence automatic detection and correction of ASR errors has become an important research area, not only for improving speech recognition accuracy but also for avoiding the propagation of the errors to the post recognition process (e.g. Machine translation and Human-Computer interaction). The aim is to be able to automatically detect, classify, and then partially or fully correct errors, regardless of the ASR system used. This can be very effective, and particularly when the ASR system is used as a black-box and the user does not have access to tune the features, the models or the decoder of the ASR system.

In the present paper we present an overview about ASR errors and the stat-of-the-art techniques for their detection and correction so as to provide a technological perspective and an appreciation of the fundamental progress that has been made in this field.

2. ASR evaluation

The performance of any ASR system is evaluated in function of the error rate. The aim of ASR evaluation is to provide a comparison criterion between different systems or techniques and to measure the performance and the progress on specific tasks based on errors statistics. There are two key areas related to ASR errors, the first one is the reference-recognised alignment which consist of finding the best word alignment between the reference and the automatic transcription and the second one is the evaluation metrics measuring the performance of the ASR systems.

2.1. Performance Factors

ASR performance is dependent upon many different factors that could be grouped in the following categories:

- **Speaker Variabilities**: Usually the acoustic model is obtained using a limited amount of speech data that characterizes the speakers at a given time and situation. However, the voice can change in time due to aging, illness, emotions, tiredness and potentially other factors. For these reasons, the acoustic model may not be representative of all speakers in all their potential states. Variabilities may not all be covered, which affect negatively the performance of the ASR systems.
- **Spoken Language Variabilities**: The spontaneous and accented speech and the high degree of pronunciation variation due to dialects, and co-articulation are known to be critical for ASR. Also, with large vocabulary, it becomes increasingly harder to find sufficient data to train the language models. Thus, subwords models are usually used instead of words models which severely degrade the performance of the recognition.
- **Mismatch Factors**: The mismatch in recording conditions between the training and testing is the main challenge for speech recognition, specially when the speech signal is acquired on telephone lines. Differences in the background noise, in the telephone handset, in the transmission channel and in the recording devices can, indeed, introduce variabilities over the recording and decrease the accuracy of the system.

2.2. Reference-Recognised Word Sequences Alignment

There are three types of errors that occur in speech recognition. First, Substitution; where a word in the reference word sequence is transcribed as a different word. Second, Deletion; where a word in the reference is completely missed in the automatic transcription. And finally, Insertion; where a word appears in the automatic transcription that has no correspondent in the reference word sequence.