A Grapheme and Phone Rescoring Combination System for Malay Broadcast News Recognition

Zainab A. Khalaf [1,2], **Tien-Ping Tan**[1], **Li-Pei Wong**[1]
[1]School of Computer Sciences, Universiti Sains Malaysia (USM)
11800 Penang, Malaysia
[2]School of Computer Sciences, Basra University
Iraq
zainab_ali2004@yahoo.com, tienping@usm.my, lpwong@cs.usm.my

***Abstract*— The main motivation of this paper is to improve the automatic speech recognition (ASR) hypothesis in the Malay language. Manual news transcription is too expensive and takes a long time. Hence, without an ASR system, access to audio archives and searches within them would be restricted to the limited number of textual documents that have been manually transcribed by humans or indexed with keywords. Multiple hypotheses are useful because the single best recognition output still has numerous errors, even for state-of-the-art systems. In this paper, we propose an approach to reduce the word error rate (WER) in an ASR hypothesis. This approach is known as the three-pass combination method using parallel ASR systems. The three-pass combination system based on grapheme rescoring and phone rescoring re-evaluates all of the hypotheses produced by the ASR systems to produce a more accurate hypothesis. To evaluate the performance of the proposed approach, Malay broadcast news contains speech from newscaster, reporter and interviewers in noisy environments recorded from Malaysia local news channels are employed. This approach reduced the WER by 4.4% from 34.5% to 30.1%. The performance of the proposed approach was compared with six approaches that are frequently used for ASR rescoring and combination.**

Index Terms— Automatic Speech Recognition, Language Model, Broadcast News, Malay Language, ASR Combination.

I. INTRODUCTION

Broadcast news keep viewers informed about the latest developments, events and issues occurring in the world. Nowadays, broadcast news can be easily accessed online. There is a rapid growth in the amount of news broadcasted from the traditional mass media such as radio, television, and cable television that are made available on the Internet. Besides that, with the availability of mobile phones with good camera, it has allowed users to record interesting videos and shared them with everyone. Now more than before, there is a need for systems capable of accessing and searching the contents of the broadcast news effectively and quickly. To allow the searching for the spoken contents in broadcast news, the spoken contents have to be first converted to text.

The speech decoding process can be implemented based on single-pass or multi-pass combination system. ASR output is typically a single-pass best hypothesis. Although automatic speech recognition is commonly used to decode speech file to text, it still produces substantial errors due to several factors that influence ASR result such as data quantity and environmental conditions [1, 2]. Multi-pass combination system is used to take the advantage of different ASR system hypotheses to find an accurate result. In a multi-pass search strategy, the first pass decodes speech to text. In the subsequent pass (or passes), rescoring is carried out using different knowledge sources to find the most likely output [3-5].

In this paper, we propose a combination approach, i.e., a three-pass combination approach. The three-pass combination system is based on grapheme rescoring and phone rescoring and it re-evaluates all of the hypotheses produced by the ASR systems that are combined during rescoring. The paper is organised as follows: related work is reviewed in Section II. Section III describes the proposed system. In Section IV, we present our experimental setup and results. Finally, conclusions are given in Section V.

II. RELATED WORK

Numerous combination systems combine the word hypotheses generated by different systems to improve ASR hypothesis [6], such as 1-best hypotheses combination, confusion network combination, *n*-best combination, lattice word combination and feature combination. The premise of combining multiple speech recognition hypothesis spaces was independently developed for lattices by [7] and for *n*-best lists by [8]. The recognizer output voting error reduction (ROVER) system combines 1-best hypotheses using voting or confidence scores [9].

The general ASR system combination principle involves the use of complementary ASR systems that exchange information at different levels of the decoding process.