Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 45 (2005) 455–470
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Confidence measures for speech recognition: A survey

Hui Jiang *

Department of Computer Science, York University, 4700 Keele Street, Toronto, Ont., Canada M3J 1P3

Received 3 August 2004; received in revised form 26 November 2004; accepted 27 December 2004

Abstract

In speech recognition, confidence measures (CM) are used to evaluate reliability of recognition results. A good confidence measure can largely benefit speech recognition systems in many practical applications. In this survey, I summarize most research works related to confidence measures which have been done during the past 10–12 years. I will present all these approaches as three major categories, namely CM as a combination of predictor features, CM as a posterior probability, and CM as utterance verification. Then, I also introduce some recent advances in the area. Moreover, I will discuss capabilities and limitations of the current CM techniques and generally comment on today's CM approaches. Based on the discussion, I will conclude the paper with some clues for future works.

Keywords: Automatic speech recognition (ASR); Confidence measures (CM); Word posterior probability; Utterance verification; Likelihood ratio testing (LRT); Bayes factors

1. Introduction

Automatic speech recognition (ASR) has achieved some substantial successes in past few decades mostly attributing to two prevalent technologies in the field, namely hidden Markov modeling (HMM) of speech signals and efficient dynamic programming search (also known as *decoding*) techniques for very-large-scale networks. Today, in many aspects, it has become a standard routine to build a state-of-the-art speech recognition system for any particular task if sufficient training data is provided for the target domain. However, when we migrate speech recognition systems from laboratory demonstrations to real-world applications, even the best ASR systems available today still encounter some serious difficulties. First of all, system performance usually dramatically degrades in the real fields because of ambient noises, speaker variations, channel distortions and many other mismatches. How to maintain and/or improve ASR performance in real-field conditions has been extensively studied in speech community under the topic of *robust*

* Tel.: +1 416 736 2100x33346; fax: +1 416 736 5872.
E-mail address: hj@cs.yorku.ca

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2004.12.004

speech recognition. Many good tutorial and overview papers, such as Juang (1991), Gong (1995), Lee (1998b) and many others, can be easily found in the literature with regard to this topic. Secondly, since every speech recognizer inevitably will make some mistakes during recognition, outputs from any ASR system are always fraught with a variety of errors. Thus, in any real-world application, it is extremely important to be able to make an appropriate and reliable judgement based on the error-prone ASR results. This requires the ASR systems to automatically assess reliability or probability of correctness for every decision made by the systems. Nowadays, to certain degree, the capability to evaluate reliability of speech recognition results has been regarded as a crucial technique to increase usefulness and "intelligence" of an ASR system in many practical applications. In this area, researchers have proposed to compute a score (preferably between 0 and 1), called *confidence measure* (CM), to indicate reliability of any recognition decision made by ASR systems. For example, a CM can be computed for every recognized word to indicate how likely it is correctly recognized or for an utterance to indicate how much we can trust the results for the utterance as a whole. Despite a large amount of research efforts in the past, we still believe that robust speech recognition and confidence measure will remain as two most active and influential research topics in speech community for a foreseeable future. Due to importance of CM in ASR systems, it has attracted considerable research attention from most major speech research groups all over the world and an excessive amount of research works have been reported in the past decade. But, unlike robust speech recognition, so far we have not seen too many overview papers in the literature to survey this important and active topic. This largely motivates me to write a comprehensive survey to summarize the CM-related research works reported mostly in the past 10–12 years. In the survey, I will mainly highlight the major progresses we have achieved in the CM area during the past decade. And I will stress some promising CM computation approaches which are theoretically sound and experimentally superior, and also discuss their capabilities and limitations. Finally, I will present some comparative discussions with respect to all reported CM computation methods and conclude the paper with some clues for possible future works from my personal perspective. Throughout the paper, I will attempt to present the CM techniques from a fairly high level and avoid technical and experimental details as much as possible, for which readers may wish to refer to the original papers. At the end of this paper, I also compose a comprehensive list of reference papers for the convenience of readers, which includes most of published works relevant to confidence measures in ASR. To my best knowledge, Lee (2001) seems to be the only CM-related overview paper which gives some good tutorials on statistical nature of confidence measure problems and also enumerates many potential CM applications for ASR.

First of all, we can backtrack some early research works on confidence measure (CM) to non-keyword rejection in word-spotting systems which were proposed to handle unconstrained speech inputs, such as Wilpon et al. (1990), Mathan and Miclet (1991), Chigier (1992), Rose (1992), Sukkar and Wilpon (1993), etc. In these works, they first adopted the so-called *garbage* or *sink* models to explicitly model non-keywords, extraneous speech and background noises in unconstrained input utterances, with which keyword spotting systems first recognize speech inputs to detect all embedded keywords as well as other speech segments corresponding to non-keywords or noises. Besides all of these, they all noticed a need to build additional rejection module to effectively distinguish non-keywords from the detected keywords in order to reduce false alarms in non-keyword rejection. Apparently, the rejection module can be viewed as a stage to investigate reliability or confidence measures for the decisions made by word-spotters. Secondly, other early CM-related works lie in automatic detection of new words (out of the current lexicon) in large vocabulary speech recognition, such as Asadi et al. (1990), Young and Ward (1993) and Young (1994), etc. In addition to modeling out-of-vocabulary (OOV) words with a (or a set of) generic hidden Markov model(s), Young and Ward (1993) proposed to use word score normalization to detect misrecognition and out-of-vocabulary words