Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 54 (2012) 1014–1028

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Joint estimation of confidence and error causes in speech recognition

Atsunori Ogawa [*], Atsushi Nakamura

NTT Communication Science Laboratories, NTT Corporation, 2–4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan

Received 2 November 2011; received in revised form 9 March 2012; accepted 17 April 2012
Available online 18 May 2012

Abstract

Speech recognition errors are essentially unavoidable under the severe conditions of real fields, and so *confidence estimation*, which scores the reliability of a recognition result, plays a critical role in the development of speech recognition based real-field application systems. However, if we are to develop an application system that provides a high-quality service, in addition to achieving accurate confidence estimation, we also need to extract and exploit further supplementary information from a speech recognition engine. As a first step in this direction, in this paper, we propose a method for estimating the confidence of a recognition result while *jointly detecting the causes of recognition errors* based on a discriminative model. The confidence of a recognition result and the nonexistence/existence of error causes are naturally correlated. By directly capturing these correlations between the confidence and error causes, the proposed method enhances its estimation performance for the confidence and each error cause complementarily. In the initial speech recognition experiments, the proposed method provided higher confidence estimation accuracy than a discriminative model based state-of-the-art confidence estimation method. Moreover, the effective estimation mechanism of the proposed method was confirmed by the detailed analyses.

Keywords: Speech recognition; Confidence estimation; Error cause detection; Joint estimation; Discriminative model

1. Introduction

Recently, speech recognition based application systems have been developed that assume the users to be the general public. These systems include voice search services and speech-to-speech translation services, and the techniques they employ have attracted keen interest at major academic conferences, e.g. (ICASSP Special Session (SS-6), 2008; Interspeech Special Session (Wed-Ses-S1), 2009; Interspeech Special Highlight Session, 2010). Not only companies but also individual developers can easily develop these application systems by using open application program interfaces (APIs), e.g. (Google Inc., 2011; Microsoft Corporation, 2011). However, since these real-field services are used by various users in various environments, it is difficult for speech recognition engines to consistently provide sufficiently accurate recognition results.

To maintain and/or improve speech recognition performance even under the severe conditions of real fields, considerable effort has been made to ensure robustness in speech recognition (e.g. Lee, 2001 for a survey). Since the basic function of a speech recognition engine is to decode a speech signal into a corresponding sequence of words, “robustness” here means that how it can output an accurate word sequence even under severe conditions. In addition to ensuring robustness, *confidence estimation* is critical. Since recognition errors are essentially unavoidable under severe conditions, as shown in Fig. 1, a recognition engine is needed to quantify the confidence in each recognized word sequence, and to output it as supplementary information, so that an application program can deal with each word sequence with its reliability. Therefore, many studies have been undertaken with the aim of improving accuracy in confidence estimation (e.g. Jiang, 2005 for a survey).

[*] Corresponding author. Tel.: +81 774 93 5358; fax: +81 774 93 1945.
E-mail addresses: ogawa.atsunori@lab.ntt.co.jp (A. Ogawa), nakamura.atsushi@lab.ntt.co.jp (A. Nakamura).

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.04.004

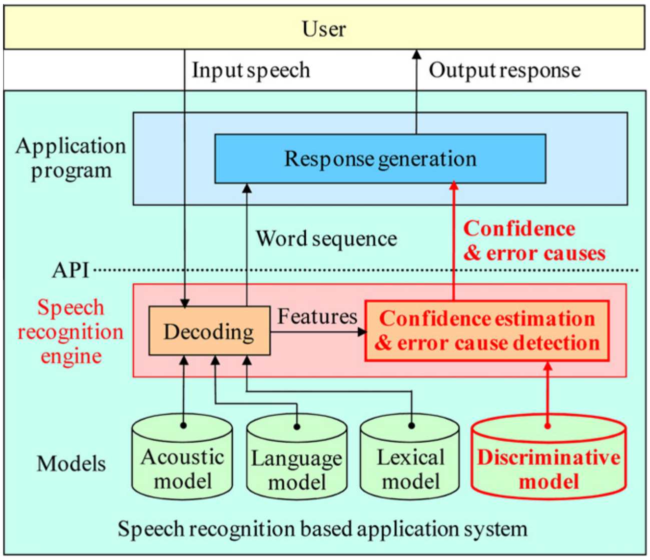

Fig. 1. Speech recognition engine that outputs causes of recognition errors together with confidence as supplementary information.

Although methods for ensuring robustness and accurate confidence estimation have been actively studied and their performance steadily improved, current speech recognition based application systems still have many points that must be improved if they are to provide the services that satisfy many users, e.g. (Schalkwyk et al., 2010; Zhou et al., 2011). In one recent report (Wang et al., 2008), many voice search dialog systems have automation rates around or below 50% in field trials. Common reasons for this are the insufficient level of robustness (Zhou et al., 2011) and the insufficient level of confidence estimation performance (Schalkwyk et al., 2010; Wang et al., 2008) of the speech recognition engines. Other reasons include the fact that outputs from recognition engines do not contain sufficient information to allow system developers to make good use of the full potential of the recognition engines (Nakano et al., 2007). For example, to obtain sufficient speech recognition accuracy for a given service, a system developer must appropriately select models associated with a recognition engine, e.g. an acoustic model, a language model, a lexical model (a word pronunciation dictionary), (Fig. 1). At the same time, the developer must take account of the speech recognition task to be executed in the service being developed, as well as the conditions under which the service will be used, e.g. various users (speakers) and acoustical environments. A developer also has to be careful when setting the parameters of a recognition engine, e.g. the beam width for pruning unlikely candidates during decoding and the scaling factor for a language prior probability. It is commonly known to experts in speech recognition technology that such selection and settings, i.e. adjustments, are crucial to maintaining recognition speed and accuracy, and that they have to be undertaken in conjunction with observations of recognition engine behavior. However, since many system developers are not experts in speech recognition technology (Nakano et al., 2007), it is not easy for them to imagine what is happening in a running recognition engine. As a result, adjustments often fail or are omitted and the engine provides suboptimal performance. Some kind of supplementary information from the engine should help developers to make the proper adjustments. Clearly confidence is less useful for this purpose.

To overcome this problem, we have been studying ways of enhancing the functions of a speech recognition engine so that it can output supplementary information in addition to confidence that is useful when employing a recognition engine as part of a real-field application system. In this paper, we focus on “what has caused the recognition error”, as such a type of supplementary information, and propose a method for detecting the *causes of recognition errors* together with the confidence (Ogawa and Nakamura, 2009, 2010a,b) (Fig. 1). We here assume, for example, Out-Of-Vocabulary (OOV) utterances and the interference of noises over speech, as the possible causes of recognition errors. If these types of information are accurately provided by a recognition engine, they can greatly help a developer to set up, adjust and refine the recognition engine to make it suitable for a specific application. For example, if it comes to a developer’s attention that users utter OOV words, he/she can add those OOV words to the word pronunciation dictionary. Also if a developer knows that input speech suffers greatly from noise interference, he/she can replace the acoustic model with one that covers such an acoustic environment or perhaps design a flow in an application program that asks the users to move to a quieter place and speak there.

We define our proposal, namely the *joint estimation of the confidence and error causes*, as a classification problem based on high-dimensional features that are obtained along with a recognized word sequence by the recognition process of a speech recognition engine. Recently proposed confidence estimation methods that exhibit good performance, e.g. (Fayolle et al., 2010; White et al., 2007), make use of many features from a recognized word sequence as with conventional confidence estimation methods (Jiang, 2005), and score the reliability of the recognized word sequence by using discriminative models, e.g. a maximum entropy model (Berger et al., 1996) and a conditional random field (Lafferty et al., 2001). As an extension of the confidence estimation methods based on discriminative models, we formulate our proposed method for jointly estimating confidence and error causes (Fig. 1).

OOV word utterances can never be correctly decoded and thus their detection is the most important type of error cause estimation. As with confidence estimation, there have been many studies on OOV word utterance detection, e.g. (Burget et al., 2008; White et al., 2008). These studies mainly focused on finding effective features for OOV word detection, and such features are naturally also effective for confidence estimation. However, in these studies, confidence estimation and OOV word detection were conducted separately. Or in (Hazen and Bazzi, 2001), OOV word