Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 28 (2014) 1287–1297

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Efficient data selection for speech recognition based on prior confidence estimation using speech and monophone models☆

Satoshi Kobashikawa*, Taichi Asami, Yoshikazu Yamaguchi, Hirokazu Masataki, Satoshi Takahashi

NTT Media Intelligence Laboratories, NTT Corporation, 1-1 Hikari-no-oka, Yokosuka-Shi, Kanagawa 239-0847, Japan

Received 22 March 2013; received in revised form 5 August 2013; accepted 5 May 2014
Available online 13 May 2014

Abstract

This paper proposes an efficient speech data selection technique that can identify those data that will be well recognized. Conventional confidence measure techniques can also identify well-recognized speech data. However, those techniques require a lot of computation time for speech recognition processing to estimate confidence scores. Speech data with low confidence should not go through the time-consuming recognition process since they will yield erroneous spoken documents that will eventually be rejected. The proposed technique can select the speech data that will be acceptable for speech recognition applications. It rapidly selects speech data with high prior confidence based on acoustic likelihood values and using only speech and monophone models. Experiments show that the proposed confidence estimation technique is over 50 times faster than the conventional posterior confidence measure while providing equivalent data selection performance for speech recognition and spoken document retrieval.

MSC: 68T10

PACS: 43.72.Ne

Keywords: Speech recognition; Spoken document retrieval; Data selection; Context independent model; Gaussian mixture model

1. Introduction

Massive quantities of videos and dialogs are stored every day; typical examples are video sharing services on the Internet and call center services provided by companies. Speech recognition technologies can transcribe the spoken components of these items automatically thus making the items searchable via their transcripts (Albertia et al., 2009). Several studies have analyzed customer needs by employing text mining (Subramaniam et al., 2009; Garnier-Rizet et al., 2008) and extracting the reasons for the calls (Fukutomi et al., 2011) from stored conversational spoken documents. A typical call center will store several tens of thousands of calls per day, and we believe that not all calls should be transcribed for the following three reasons. (1) The computation cost involved in transcribing all calls is excessive. (2)

☆ This paper has been recommended for acceptance by 'Prof. R.K. Moore'.
* Corresponding author.

http://dx.doi.org/10.1016/j.csl.2014.05.001
0885-2308/

An informative analysis can be achieved from a subset of the calls. (3) The quality of the recorded speech samples varies (Benzeghiba et al., 2007), and erroneous speech recognition (due to the poor input) will degrade the efficiency of subsequent spoken document retrieval (Sanderson and Shou, 2007) and analysis.

Several confidence measures have been proposed for identifying "accurate" speech samples (Jiang, 2005). Unfortunately, they require the computationally expensive step of speech recognition processing to obtain confidence scores, which are estimated from the recognition results; they waste considerable computer resources on samples that will eventually be rejected. Most conventional methods target word or utterance verification. A dialog (similar to spoken document) level confidence measure has been proposed (Litman et al., 1999), but it is also computationally inefficient because it requires several features including speech recognition results to estimate confidence. Several data selection methods have been proposed (Wu et al., 2007), but their target is to select training data, so they fail to reduce the computation cost significantly.

Our proposal efficiently identifies speech samples that will be well recognized with an extremely low computation cost prior to speech recognition. It can identify those samples that have high confidence levels from massive numbers of stored speech samples. Prior confidence must be estimated rapidly because speech recognition can only proceed after the estimation results have been received. The proposed estimation technique utilizes the acoustic model used for posterior speech recognition. The proposal uses only context independent (monophone) models and speech models to reduce the computation cost. For even greater efficiency, its confidence estimation step eliminates all processing other than the calculation of acoustic output likelihood from Gaussian mixture models (GMMs). The prior confidence is calculated frame by frame from the difference between the output log-likelihoods of the monophone and speech GMMs. This confidence formulation is an approximation of the state level posterior probability with the state occurrence probability. This paper evaluates the actual efficiency of our technique in speech recognition and spoken document retrieval tasks. Experiments show that the proposed technique is significantly faster than the conventional posterior confidence measure based on speech recognition, while maintaining equivalent data selection performance.

The rest of this paper is organized as follows. Related work is outlined in Section 2. The proposed technique is described in Section 3. Section 4 introduces experiments conducted to confirm the effectiveness of the proposed technique. Our conclusion is presented in Section 5.

2. Related work on data selection for speech recognition and its application

Since there are many factors that cause variability in speech signals (Benzeghiba et al., 2007), the recognition accuracy is strongly dependent on the data. Several data selection methods have been proposed for training (Wu et al., 2007; Lin and Bilmes, 2009) and adapting (Cincarek et al., 2006) acoustic models for speech recognition. Wu et al. also selected data to be transcribed for training by using the confidence score (Wu et al., 2011); this technique is called active learning. A great number of confidence measure methods have been proposed (Jiang, 2005) and they could also be useful for selecting data during speech recognition processing, since inaccurately recognized data impacts negatively on the subsequent application. Stoyanchev et al. detected misrecognized words in spoken dialog systems (Stoyanchev et al., 2012). Seigel et al. estimated a confidence measure at the word/utterance level by using conditional random fields (CRF). Ogawa et al. also used CRF directly to estimate the recognition rate rather than the confidence score both per utterance and per lecture at the spoken document level (Ogawa et al., 2012). Asami et al. also estimated the spoken document confidence score by using contextual coherence (Asami et al., 2011). Senay et al. detected low-quality documents by using a confidence measure and semantic consistency based on the latent Dirichlet allocation (LDA) model for spoken document retrieval (Senay and Linarès, 2012). Li et al. used semantic similarity to estimate a confidence measure for spoken term detection (Li et al., 2012). There are several confidence measure methods at a variety of levels depending on the application.

Conventional confidence measure estimations require speech recognition results; this means that a lot of computation time is required to recognize low-confidence and unuseful data, which should be rejected. Thus, we attempt to reject unuseful data at the document level to prevent harmful effects on the subsequent application prior to speech recognition. In a conventional approach, Lee et al. proposed rejecting data before speech recognition by using noise GMMs (Lee et al., 2004). However, this method could reject data at the utterance level and needs to know the noise type beforehand. Chang et al. also proposed a pre-rejection algorithm that enhances the robustness of speech recognition by using pitch correlation (Seo et al., 2003), which allows it reject seriously distorted speech signal during wireless communication. However, it fails to reject slightly distorted speech with pitch continuity. This paper proposes an efficient method for