Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 36 (2016) 24–41

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Differenced maximum mutual information criterion for robust unsupervised acoustic model adaptation☆

Marc Delcroix*, Atsunori Ogawa, Seong-Jun Hahm[1], Tomohiro Nakatani, Atsushi Nakamura[2]

NTT Communication Science Laboratories, NTT Corporation, 2-4, Hikaridai Seika-cho, Souraku-gun, Kyoto 619-0237, Japan

Received 23 October 2014; received in revised form 23 July 2015; accepted 3 August 2015
Available online 19 August 2015

Abstract

Discriminative criteria have been widely used for training acoustic models for automatic speech recognition (ASR). Many discriminative criteria have been proposed including maximum mutual information (MMI), minimum phone error (MPE), and boosted MMI (BMMI). Discriminative training is known to provide significant performance gains over conventional maximum-likelihood (ML) training. However, as discriminative criteria aim at direct minimization of the classification error, they strongly rely on having accurate reference labels. Errors in the reference labels directly affect the performance. Recently, the differenced MMI (dMMI) criterion has been proposed for generalizing conventional criteria such as BMMI and MPE. dMMI can approach BMMI or MPE if its hyper-parameters are properly set. Moreover, dMMI introduces intermediate criteria that can be interpreted as smoothed versions of BMMI or MPE. These smoothed criteria are robust to errors in the reference labels. In this paper, we demonstrate the effect of dMMI on unsupervised speaker adaptation where the reference labels are estimated from a first recognition pass and thus inevitably contain errors. In particular, we introduce dMMI-based linear regression (dMMI-LR) adaptation and demonstrate significant gains in performance compared with MLLR and BMMI-LR in two large vocabulary lecture recognition tasks.

Keywords: Discriminative criterion; Differenced maximum mutual information; Speech recognition; Acoustic model adaptation; Unsupervised adaptation

1. Introduction

Discriminative criteria have been widely used for automatic speech recognition (ASR) for training acoustic models (Nadas et al., 1988; Povey and Woodland, 2002; McDermott et al., 2007; Heigold et al., 2012), language models (Kuo et al., 2002) or feature transforms (Povey et al., 2005, 2008; Droppo and Acero, 2005; Zhang et al., 2006). Discriminative

☆ This paper has been recommended for acceptance by Mark J.F. Gales.

* Corresponding author at: NTT Communication Science Laboratories, Media Information Laboratory, 2-4, Hikaridai, Seika-cho, Keihanna Science city, Kyoto 619-0237, Japan. Tel.: +81 774 93 5288; fax: +81 774 93 5158.
E-mail address: marc.delcroix@lab.ntt.co.jp (M. Delcroix).
[1] Now with the University of Texas at Dallas (UTD), United States.
[2] Now with the Graduate School of Natural Sciences, Nagoya City University, Japan.

http://dx.doi.org/10.1016/j.csl.2015.08.001
0885-2308/

criteria aim at the direct minimization of classification error. Therefore, they are better correlated to word error than conventional maximum likelihood (ML). Consequently, discriminative training has provided significant performance improvement for many tasks and has became the de facto procedure for training ASR systems.

Many discriminative criteria have been proposed, including maximum mutual information (MMI) (Nadas et al., 1988), minimum phone error (MPE) (Povey and Woodland, 2002), and minimum classification error (MCE) (Juang and Katagiri, 1992). More recently, margin-based extensions of these criteria have also been introduced such as soft margin (Li et al., 2006), boosted MMI (BMMI) (Povey et al., 2008) and boosted MPE (Heigold et al., 2008). These various discriminative criteria differ in their formulations, but they share the same principle of aiming at increasing the classification scores for the reference labels, while decreasing the scores for competing recognition hypotheses. As a consequence, *discriminative approaches usually require to have high quality reference labels* to work properly.

In many situations, it is difficult or costly to obtain reference labels without transcription errors. For example, transcribing large amounts of training speech data is not only very expensive but also difficult, especially when dealing with spontaneous speech. Semi-supervised training is a practical approach for addressing this issue. With semi-supervised training, only a part of the training data is transcribed manually, and labels for the rest of the training data are generated automatically using a pre-existing recognizer (Lamel et al., 2002; Wessel and Ney, 2005; Wang et al., 2007). A similar approach is used for the unsupervised adaptation of acoustic models where labels also need to be estimated automatically from the untranscribed adaptation data. In such cases, the *estimated reference labels inevitably contain recognition errors*. Therefore, using a conventional discriminative criterion in such situations is particularly challenging (Wang et al., 2007).

In this paper, we investigate the use of the recently proposed differenced MMI (dMMI) (McDermott et al., 2009, 2010) criterion in the presence of errors in the reference labels. We argue that dMMI is well suited for such situations. dMMI was first introduced as a generalization of existing criteria such as MPE and MMI. dMMI has been successfully used for training acoustic models (McDermott et al., 2010), discriminative feature transforms (Delcroix et al., 2012) and the weights of a weighted finite state transducer (WFST) used for ASR decoding (Kubo et al., 2012, 2012, 2013). dMMI can be derived as the integration over a margin interval of margin-based MPE objective functions. It can also be interpreted as a smoothed version of the BMMI objective function. With dMMI, the reference labels can be defined in *a smoothed manner*, i.e. as a summation of the contribution of each recognition candidate weighted by margin terms that emphasize the contribution of candidates close to the reference labels. In other words, the numerator of the dMMI objective function does not consist only of the contribution of the reference label, but also considers recognition candidates close to that reference label. Therefore, *dMMI has an intrinsic mechanism for mitigating the influence of errors in the reference labels*.

We demonstrate the robustness of dMMI to errors in the reference labels for unsupervised acoustic model adaptation. Acoustic models must be adapted to mitigate the mismatch that often occurs between training and testing conditions due to unseen speaker or acoustic conditions (Yoshioka et al., 2014). In many cases, the adaptation has to be performed in an unsupervised way because transcribed data may not always be available. There is a great interest in using a discriminative criterion for adaptation. Indeed, discriminative criteria can potentially improve performance compared with the ML criterion and can better preserve the discriminative capabilities of discriminatively trained acoustic models. Therefore, discriminative unsupervised adaptation is an important application in itself. Moreover, it is a good example of an application where the reference labels contain errors and therefore is directly related to semi-supervised training.

There have been many investigations of discriminative adaptation (Gunawardana and Byrne, 2001; Uebel and Woodland, 2001; Povey et al., 2003; Chien and Huang, 2006; Wu and Huo, 2007; Wang and Woodland, 2008; Matsuda et al., 2009; Gibson and Hain, 2012). Most approaches are based on the maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995; Gales and Woodland, 1996) adaptation, but propose replacing the ML criterion with a discriminative one. Most studies target supervised adaptation. However, some have recognized the challenge of unsupervised discriminative adaptation and proposed approaches for handling the problem caused by inaccurate reference labels (Wang and Woodland, 2008; Gibson and Hain, 2012; Yu et al., 2009). For example, Wang and Woodland (2008) and Gibson and Hain (2012) proposed focusing on adaptation data that are expected to be correctly transcribed. This was achieved by weighting the MPE objective function with a word/phoneme correctness estimation.

In this paper, we introduce dMMI-based LR adaptation. We demonstrate both theoretically and experimentally that *dMMI-LR is well suited for unsupervised adaptation*. Indeed, the smoothed reference definition of dMMI can achieve a similar effect to Wang and Woodland (2008) and Gibson and Hain (2012) without the need for an explicit estimation of the word/phoneme correctness. This paper is an extension of our previous work (Delcroix et al., 2013) and includes