ARTICLE IN PRESS
JID: YCSLA [m3+;November 16, 2016;19:51]
Available online at www.sciencedirect.com
ScienceDirect
International Journal of Impact Engineering xxx (2016) xxx-xxx
www.elsevier.com/locate/ijimpeng

Sparse coding over redundant dictionaries for fast adaptation of speech recognition system

Q1 S. Shahnawazuddin*, Rohit Sinha

Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India

Received 23 April 2015; received in revised form 4 April 2016; accepted 28 October 2016

Abstract

This work presents a novel use of the sparse coding over redundant dictionary for fast adaptation of the acoustic models in the hidden Markov model-based automatic speech recognition (ASR) systems. The presented work is an extension of the existing acoustic model-interpolation-based fast adaptation approaches. In these methods, the basis (model) weights are estimated using an iterative procedure employing the maximum-likelihood (ML) criterion. For effective adaptation, typically a number of bases are selected and as a result of that the latency of the iterative weight estimation process becomes high for those ASR tasks that involve human-machine interactions. To address this issue, we propose the use of sparse coding of the target mean supervector over a speaker-specific (exemplar) redundant dictionary. In this approach, the employed greedy sparse coding not only selects the desired bases but also compresses them into a single supervector, which is then ML scaled to yield the adapted mean parameters. Thus reducing the latency in the basis weight estimation in comparison to the existing fast adaptation techniques. Further, to address the loss in information due to reduced degrees of freedom, we have also extended the proposed approach using separate sparse codings over multiple (exemplar and learned) redundant dictionaries. In adapting an ASR task involving human-computer interactions, the proposed approach is found to be as effective as the existing techniques but with a substantial reduction in the computational cost.

Keywords: Fast adaptation*; Acoustic model interpolation; Sparse coding; Exemplar and learned speaker dictionary*

1. Introduction

The automatic speech recognition (ASR) systems are traditionally developed using the Gaussian mixture model (GMM) based context-dependent hidden Markov model (CD-HMM). Recently, the deep neural network (DNN) (Hinton et al., 2012) has been proposed to overcome the inefficiency of the GMM in the modeling of the events that lie on or near a nonlinear manifold in the data. With the advent of the DNN for generating the observation probabilities, the ASR systems based on DNN−HMM are fast becoming popular[1]. In general, the ASR systems are trained on speech data from a large number of speakers (male and female). This pooled training is primarily intended towards

* Corresponding author.
E-mail address: s.syed@iitg.ernet.in (S. Shahnawazuddin), rsinha@iitg.ernet.in (R. Sinha).

[1] Even though, the work presented in this paper, while being targeted to the GMM−HMM systems, remains relevant as some of the classical GMM−HMM adaptation approaches have been ported to the DNN-based systems.

http://dx.doi.org/10.1016/j.csl.2016.10.004
0885-2308/ 2016 Elsevier Ltd. All rights reserved.

making the statistical models speaker independent (SI). Unlike an speaker dependent (SD) system, the SI systems have to deal with both the intra-speaker and the inter-speaker variability. As a result of that, an SI system is reported to be 2 to 3 times inferior in comparison to an SD system when both the systems are trained with an equal amount of data (Woodland, 2001). The SD systems, though quite effective, are infeasible to be built for each of the speakers in the user population on account of requiring a large amount of speech data per speaker. Consequently, the speaker adaptation techniques have been developed which intend to modify the parameters of the SI system to better suit a particular speaker given the limited amount of data from that speaker.

Some of the applications of ASR are information retrieval, language learning tools, voice-based search and entertainment (Wang et al., 2008; Eskenazi, 2009; Schalkwyk et al., 2010; Gray et al., 2014). In the case of ASR tasks involving human-machine interactions, the system is required to recognize speech from the adult as well as the child (male/female) speakers. Further in such ASR tasks, the adaptation data is made available to the ASR system in an incremental manner. Moreover, the duration of the data is generally very small. Consequently, the conventional adaptation techniques like the maximum *a-posteriori* (MAP) (Gauvain and Lee, 1994) and the maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995; Digalakis et al., 1995) are found to be unsuitable. This is so because the available adaptation data is insufficient to estimate a large number of parameters. On the other hand, the fast adaptation techniques try to modify the model parameters even with a small amount of adaptation data (Hazen and Glass, 1997; Gales, 1999; Kuhn et al., 2000; Kenny et al., 2005). In these techniques the adapted model parameters are derived by a linear interpolation of a set of bases (predefined acoustic models) spanning a low (K) dimensional subspace. Consequently, only a few interpolation weights (or the direction coordinates) need to be estimated. Such a reduction in complexity makes these approaches amenable for aforementioned adaptation tasks. The fast adaptation approaches are found to be effective for the adaptation of the Gaussian means as well as the mixture-weights (Duchateau et al., 2008; Hahm et al., 2010) of the acoustic model.

In the initial works on fast adaptation (Hazen and Glass, 1997; Gales, 1999; Kuhn et al., 2000; Kenny et al., 2005), the K bases to be interpolated were kept fixed. The recent works (Mak et al., 2006; Teng et al., 2009) have shown that an improved recognition performance can be achieved for the given test data by selecting K bases from a predefined set of acoustic models. In those approaches, speaker adapted (SA) models corresponding to each of the N speakers in the training set are derived first. This is usually done by adapting the mean parameters of the SI model with the speaker-specific data while keeping the covariance matrices and the mixture-weights unchanged. Given the test data, a set of K models (bases) are then selected from the N acoustic models and are linearly combined. The K interpolation weights required for linear combination are estimated using an iterative maximum-likelihood (ML) approach. It is worth mentioning here that there do exist some fast adaptation approaches in which the bases are derived using the given adaptation data rather than using predefined ones (Gales, 1999; Kenny et al., 2005).

The aforementioned fast adaptation techniques differ mainly in the way the K bases are created/selected. The work presented in Mak et al. (2006) employs a Vitebi-alignment-based ML search for the basis selection. The ML search becomes very cumbersome when N is large and hence is not suitable for adapting interactive ASR tasks. The approach reported in Teng et al. (2009) requires the estimation of the interpolation weights for all the N models using a single iteration of ML estimation. The K acoustic models having the largest magnitude for the interpolation weights are selected for deriving the adapted model. The interpolation weights for those K models are then re-estimated iteratively. The complexity introduced in estimating weights for all the N models limits the feasibility of this approach in the case of ASR tasks involving human-machine interactions.

The work presented in this paper is inspired by the recent works employing a dynamic selection of acoustically close bases. In our earlier work (Shahnawazuddin and Sinha, 2014a), techniques employing the sparse representation (SR) (Elad, 2010) over a redundant dictionary were explored for the dynamic selection of the bases. In that work, the sparse coding was used only for the basis selection while the basis coefficients in the sparse coding were discarded. Like the other model-interpolation-based techniques, the interpolation weights were estimated iteratively in the ML sense. The latency in the basis selection process was reduced due to the greedy SR-based approaches but the computational cost of the weight estimation remained the same. In contrast to our earlier work, in this paper we propose a novel use of the sparse coding to derive the adapted model mean parameters.[2] The main contributions of the work are as follows:

[2] An initial version of this work was presented in *INTERSPEECH 2014* (Shahnawazuddin and Sinha, 2014b)

Please cite this article as: S. Shahnawazuddin, R. Sinha, Sparse coding over redundant dictionaries for fast adaptation of speech recognition system, International Journal of Impact Engineering (2016), http://dx.doi.org/10.1016/j.csl.2016.10.004