Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 30 (2015) 116–134

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Relevance factor of maximum *a posteriori* adaptation for GMM–NAP–SVM in speaker and language recognition☆

Chang Huai You*, Haizhou Li, Kong Aik Lee

*Institute for Infocomm Research, A*STAR, Singapore*

Received 1 December 2012; received in revised form 24 March 2014; accepted 11 September 2014
Available online 29 September 2014

Abstract

This paper studies the relevance factor in maximum *a posteriori* (MAP) adaptation of Gaussian mixture model (GMM) for speaker and language recognition. Knowing that relevance factor determines how much the observed training data influence the model adaptation, thus the resulting GMM model, it is believed that more effective modeling can be achieved if the relevance factor is adaptive to the corresponding data. We therefore provide a mathematic derivation for the estimation of relevance factor. GMM supervector support vector machine (SVM) with nuisance attribute projection (NAP) (GMM–NAP–SVM) has been reported to be effective and reliable for speaker and language recognition. Being a discriminative classifier in nature, a GMM–NAP–SVM system is sensitive to the magnitude and direction of a supervector in the high dimensional space. However, when characterizing a speech utterance with GMM supervector estimated through MAP, we observe that the resulting supervector is undesirably affected by the varying duration of the utterance. We propose an adaptive relevance factor that adapts to the duration to mitigate the variability effect due to the length of utterance. We give a systematic investigation on different types of relevance factor of MAP in different applicatively platforms. We show the efficacy of the data-dependent as well as adaptive relevance factors on the National Institute of Standards and Technology (NIST) speaker recognition evaluation (SRE) 2008 and language recognition evaluation (LRE) 2009 and 2011 tasks respectively.

Keywords: Maximum *a posteriori*; Supervector; Gaussian mixture model; Support vector machine

1. Introduction

Gaussian mixture model (GMM) that relies on acoustic spectral features has shown very effective and reliable performance for text-independent speaker and language recognition (Reynolds et al., 2000); especially GMM-supervector with the application of support vector machine (SVM) has its effective performance (Campbell et al., 2006). In GMM approach, a model is obtained by maximum *a posteriori* (MAP) estimation from a universal background model (UBM) (Gauvain and Lee, 1994). A UBM is usually trained through expectation-maximization (EM) algorithm from a

☆ This paper has been recommended for acceptance by R.K. Moore.

* Corresponding author. Tel.: +65 6408 2763; fax: +65 6776 1378.
E-mail address: echyou@i2r.a-star.edu.sg (C.H. You).

http://dx.doi.org/10.1016/j.csl.2014.09.003
0885-2308/

background data to cover a wide range of languages, speakers, sessions and channels. With MAP, we adapt the UBM towards a GMM using speech utterance(s).

A GMM-supervector carries rich information from its corresponding utterance. Besides the desired information such as speaker or language characteristics for recognition purpose, it also contains some unwanted information such as channel and duration of utterance. Such distraction may lead to the inconsistency among the involved GMM-supervectors and thus result in mismatch between the training and testing utterances. In (Kenny, 2005) (Kenny et al., 2007) (Kenny et al., 2005), a joint factor analysis (JFA) is introduced, where a GMM-supervector is viewed as a combination of different supervectors for different factors such as channel and voice factors. The JFA is used to compensate the channel variation through eigenchannel modeling in GMM-supervector and to emphasize the speaker-dependent component by using low dimension speaker factor through eigenvoice modeling. Actually, eigenvoice modeling is more useful for insufficient enrollment data. In (Kenny et al., 2003), eigenchannel MAP is proposed by using a prior distribution on channel compensations to adapt the speaker GMM to the test utterance. Presently, the i-vector technique that was originated from JFA brings a new height to speaker recognition and becomes the most popular (Dehak et al., 2009, 2011). The i-vector extractor converts a sequence of features into a single low-dimensional vector in the total variability space, by which speech segment of variable length can be represented as fixed-length vector. In this regard, linear discriminant analysis (LDA) (McLaren and van Leeuwen, 2011), probabilistic LDA (PLDA) (Prince, 2007), and the heavy-tailed PLDA (Kenny, 2010) are useful for i-vector system.

We know that, in GMM–SVM system, the nuisance attribute projection (NAP) (Solomonoff et al., 2004, 2005; Campbell et al., 2006) is effective for channel compensation. NAP realizes channel compensation by removing the nuisance directions which are indicated with the greatest variation after subtracting the effect of the particular speaker (McLaren et al., 2007). One reason that the NAP is used in SVM kernel is due to its simple implementation while achieving comparable performance to the factor analysis model (McLaren et al., 2009). In fact, GMM-supervector system can work effectively with NAP without eigenchannel and eigenvoice analysis. In this paper, we are interested in the GMM–NAP–SVM without factor analysis.

We understand that GMM–NAP–SVM concerns more with the GMM parameter than the GMM probability, it is therefore straightforward to approach the problem in the supervector domain. Because the relevance factor implicitly controls the degree of the contribution of new data to the updating of parameters (i.e., weight, mean, covariance), an estimation of relevance factor may offer a new way to optimize the effect of different background data.

With GMM–NAP–SVM, it is common that we use conventional MAP with a relevance factor that is set empirically. In this case, the relevance factor is not optimized towards a particular application. Due to the nature of generative modeling (Reynolds et al., 2000), in GMM–UBM system, the relevance factor is less sensitive and therefore can be fixed. However, SVM works in a discriminative manner, it is sensitive to the variation of GMM supervector. We have a good reason to expect that GMM–SVM can benefit from an accurate estimate of the relevance factor, that allows a supervector to manifest the saliency of target characteristics.

In this paper, we start by introducing the mathematical derivation of the data-dependent relevance factor for GMM–SVM system in connection with the universal background data, where we analyze the relevance factor of MAP through the supervector modeling and derive specifically the algorithm to obtain the relevance factor as well as the related parameters. Then, we analyze the duration effect. In Ben and Bimbot (2003), the weighting coefficients are adapted to the Kullback-Leibler (KL) distance in MAP estimation, where the symmetric KL distance is constrained to its maximum likelihood (ML) estimate. In an earlier report (Sönmez et al., 1999), a duration factor of the test segment was considered at the score level. The average of acoustic scores from the frames segmented as the target is augmented by statistics of duration via a thresholded nonlinear function to generate a score per test waveform. Scores averaged with less than a threshold size are decreased with a linear penalty. In Pelecanos et al. (2004), the variability utterance lengths were investigated through the score distribution analysis. While the prior work was focused on score level compensation, in this paper, we propose a compensation scheme for the duration variability of the utterance in the modeling domain. In conventional MAP adaptation for GMM–SVM, if a mixture component has a low probabilistic count on new data, the sufficient statistics will be de-emphasized, otherwise will be emphasized. As a result, the displacement of GMM-supervector from the UBM-supervector varies undesirably due to the unwanted variability of new data such as the duration of utterances. Unfortunately, the duration of utterances varies inevitably. Thus, for GMM–NAP–SVM system, we propose an adaptive relevance factor scheme to compensate the duration variation of the adaptation data. In You et al. (2010) (You et al., 2010), we have shown the effectiveness of the adaptation of the relevance factor to the duration of the particular utterance for speaker and language recognition.