Expert Systems with Applications 39 (2012) 4287–4291

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Maximum A Posteriori Linear Regression for language recognition

Jinchao Yang *, Xiang Zhang, Hongbin Suo, Li Lu, Jianping Zhang, Yonghong Yan

Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Keywords:
Language recognition
MLLR
MAPLR
SVM

ABSTRACT

This paper proposes the use of Maximum A Posteriori Linear Regression (MAPLR) transforms as feature for language recognition. Rather than estimating the transforms using maximum likelihood linear regression (MLLR), MAPLR inserts the priori information of the transforms in the estimation process using maximum a posteriori (MAP) as the estimation criterion to drive the transforms. By multi MAPLR adaptation each language spoken utterance is convert to one discriminative transform supervector consist of one target language transform vector and other non-target transform vectors. SVM classifiers are employed to model the discriminative MAPLR transform supervector. This system can achieve performance comparable to that obtained with state-of-the-art approaches and better than MLLR. Experiment results on 2007 NIST Language Recognition Evaluation (LRE) databases show that relative decline in EER of 4% and on mincost of 9% are obtained after the language recognition system using MAPLR instead of MLLR in 30-s tasks, and further improvement is gained combining with state-of-the-art systems. It leads to gains of 6% on EER and 11% on minDCF comparing with the performance of the only combination of the MMI system and the GMM–SVM system.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

The aim for language recognition is to determine the language spoken in a given segment of speech. PRLM and PPRLM approaches that use phonotactic information, have shown very successful performance (Yan & Barnard, 1995; Zissman, 1995). In PPRLM, several tokenizers are used to transcribe the input speech into phoneme strings or lattices (Gauvain, Messaoudi, & Schwenk, 2004; Shen, Campbell, Gleason, Reynolds, & Singer, 2006) which are scored by n-gram language models. It is generally believed that phonotactic feature and spectral feature provide complementary cues to each other (Zissman, 1995). The spectral features of speech are collected as independent vectors. The collection of vectors can be extracted as shifted-delta-cepstral acoustic features, and then modeled by Gaussian Mixture Model (GMM). The result was reported in Torres-Carrasquillo, Singer, Kohler, Greene, and Reynolds (2002). The approach was further improved by using discriminative train that named Maximum Mutual Information (MMI).

Several studies using SVM in language recognition to form GSV-SVM system (Campbell, Campbell, Reynolds, Singer, & Torres-Carrasquillo, 2006; Li, Ma, & Lee, 2007). SVM as a classifier maps input feature vector into high dimensional space then separate classes with maximum margin hyperplane. It is important to choose an appropriate SVM feature expansion, which maps a given utterance to a feature vector in a high-dimensional feature space for SVM classification.

Maximum likelihood linear regression (MLLR) is a commonly used adaptation approach in large vocabulary speech recognition systems. The concatenation of the transformation parameters can be seen as a kind of mapping from the given utterance to a high-dimensional space. MLLR and CMLLR are introduced for the task of speaker recognition in Ferras, Leung, Barras, and Gauvain (2008) and language recognition in Shen and Reynolds (2008) and Zhong and Liu (2010), which is useful for the system fusion. A system proposed in Stolcke, Ferrer, Kajarekar, Shriberg, and Venkataraman (2005) first used the MLLR transforms employed in automatic speech speaker recognition. Another system uses constrained MLLR (CMLLR) to adapt the means of a GMM UBM to a given utterance, and uses the entries of the transform as features for SVM classification (Ferras, Leung, Barras, & Gauvain, 2007). In Shen and Reynolds (2008) CMLLR is used as a feature-space implementation in language recognition. Zhong and Liu (2010) propose CMLLR supervector kernel and the system uses the entries of the transform as features for SVM classification in language recognition.

In MLLR and CMLLR, parameters are estimated with the maximum likelihood (ML) criteria, which is well known for its poor asymptotic properties and may generate unacceptable affine transformation parameters when the adaptation data is insufficient (Gales, 1998). A possible solution to this problem is to introduce some constraints on the possible values of the transformation

* Corresponding author.
E-mail addresses: yangjinchao@hccl.ioa.ac.cn (J. Yang), xzhang@hccl.ioa.ac.cn (X. Zhang), hsuo@hccl.ioa.ac.cn (H. Suo), luli@hccl.ioa.ac.cn (L. Lu), jzhang@hccl.ioa.ac.cn (J. Zhang), yonghongyan@hccl.ioa.ac.cn (Y. Yan).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.09.104

parameters. Maximum A Posteriori Linear Regression (MAPLR) (Chesta, Siohan, & Lee, 1999) is such an adaptation approach, which inserts the priori information of the transforms in the estimation process using maximum a posteriori (MAP) as the estimation criterion to drive the transformation parameters η:

$$\hat{\eta} = \arg\max_{\eta} p(\eta|\mathbf{X}, \lambda) = \arg\max_{\eta} p(\mathbf{X}|\lambda, \eta)p(\eta) \tag{1}$$

where $p(\eta)$ is the priori distribution of the parameters η, $\mathbf{X}$ is the adaptation features and λ represents the universal background model.

We believe MAPLR can generate transforms that could show better adaptation performance. In ours proposed MAPLR language recognition system each language spoken utterance is convert to feature vector. Then discriminative MAPLR transform supervector space is built from feature vector by multi MAPLR adaptation. That is one language spoken utterance is convert to one discriminative transform supervector consist of one target language transform vector and other non-target transform vectors. SVM classifiers are employed to model the discriminative MAPLR transform supervector and LDA and diagonal covariance gaussians are used as backend in language score calibration.

This paper is organized as following: In Section 2, we give a simple review of Support Vector Machines and MLLR. Section 3 shows the MAPLR technique. In Section 4, the proposed MAPLR language recognition system is presented in detail. corpora and evaluation and experimental result are given in Sections 5 and 6. Finally, we conclude in Section 7.

2. Background

2.1. Support Vector Machines

An SVM (Cristianini & Shawe-Taylor, 2000) is a two-class classifier constructed from sums of a kernel function $K(\cdot,\cdot)$:

$$f(x) = \sum_{i=1}^{N} \alpha_i t_i K(\mathbf{x}, \mathbf{x_i}) + d \tag{2}$$

where N is the number of support vectors, t_i is the ideal output, α_i is the weight for the support vector x_i, $\alpha_i > 0$ and $\sum_{i=1}^{N} \alpha_i t_i = 0$. The ideal outputs are either 1 or -1, depending upon whether the corresponding support vector belongs to class 0 or class 1. For classification, a class decision is based upon whether the value, $f(x)$, is above or below a threshold.

The kernel $K(\cdot,\cdot)$ is constrained to have certain properties (the Mercer condition), so that $K(\cdot,\cdot)$ can be expressed as

$$K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})'\phi(\mathbf{y}) \tag{3}$$

where $\phi(x)$ is a mapping from the input space (where $\mathbf{x}$ lives) to a possibly infinite dimensional SVM expansion space. We refer to the $\phi(x)$ as the SVM features.

2.2. Maximum likelihood linear regression

In maximum likelihood linear regression (MLLR), a single affine transformation is applied to the means of all the mixtures of the GMM–UBM to obtain the adapted means:

$$\hat{\mu}_m = \mathbf{A}\mu_m + \mathbf{b} = \mathbf{W}\xi_m \tag{4}$$

where μ_m and $\hat{\mu}_m$ represent the mean vector before and after adaptation, ξ_m is the extended mean vector, defined as $\xi_m = (\mu_m, 1)$. Thus, the transformation parameters can be denoted by $\eta = \{\mathbf{A}, \mathbf{b}\} = \{\mathbf{W}\}$. $\mathbf{A}$ and $\mathbf{b}$ are chose to maximize the likelihood that the utterance was generated by the adapted model.

MLLR will product new GMM supervector and Transform vector as Fig. 1 show. We will focus on Transform vector η which will be use as SVM feature in the SVM feature space in language recognition.

MLLR adaptation approach is a well-accepted approach for speaker adaptation, which reduces the mismatch between the speaker-independent model and the speaker-specific characteristics of the individual speakers by performing one or several affine matrix transformations on all mean vectors of the speaker-independent model to generate the speaker-dependent models. In Shen and Reynolds (2008) and Zhong and Liu (2010) MLLR is successfully used in language recognition. It show the coefficients in these MLLR matrices carry the information of the language-specific characteristics, which can be used as features in language recognition. In Zhang, Wang, Xiao, Zhang, and Yan (2010), we propose that MAPLR which inserts the priori information of the transforms in the estimation process to get better transformation parameters than MLLR in speaker recognition. We try to introduce MAPLR which is used in speech recognition and speaker recognition to language recognition firstly.

3. MAPLR for language recognition

3.1. Model transformation function

In MAPLR adaptation, the mean vectors of the Gaussian mixtures are also adapted using an affine transform as MLLR:

$$\hat{\mu}_m = \mathbf{A}\mu_m + \mathbf{b} = \mathbf{W}\xi_m \tag{5}$$

where μ_m and $\hat{\mu}_m$ represent the mean vector before and after adaptation, ξ_m is the extended mean vector, defined as $\xi_m = (\mu_m, 1)$. Thus, the transformation parameters can be denoted by $\eta = \{\mathbf{A}, \mathbf{b}\} = \{\mathbf{W}\}$.

In this work, we first assume that all the mean vectors share the same transform. Given some adaptation data of hypothesized language utterances, $\mathbf{X} = \{\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(T)\}$, the objective of the MAPLR for language recognition is to derive an affine transform η using a MAP estimation criterion as described by Eq. (1).

3.2. Definition of the auxiliary function

Commonly, the maximization of Eq. (1) cannot be carried out directly. The maximization problem is traditionally addressed by solving an auxiliary and simpler problem having the same solution, using the EM algorithm (Dempster, Laird, & Rubin, 1977). Let $\lambda = \{\omega_m, \mu_m, \mathbf{\Sigma}_m\}, m = 1, 2, \ldots, M$ denote the UBM, according to Chesta et al. (1999), the final auxiliary function in a GMM framework can be finally defined as following:

$$Q(\eta|\bar{\eta}) = \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_t(m) \log p(\mathbf{x}(t)|\eta, \mu_m, \mathbf{\Sigma}_m) + \log p(\eta) + \Psi \tag{6}$$

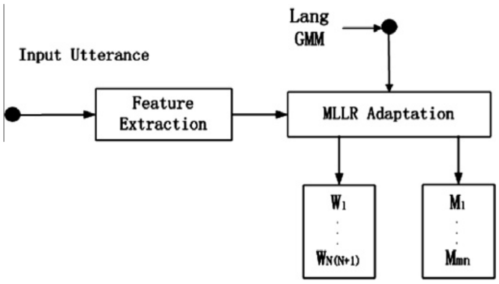

Fig. 1. Maximum likelihood linear regression adaptation flow.

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات