Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 73 – 79

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

Unsupervised Linear Discriminant Analysis for Supporting DPGMM Clustering in the Zero Resource Scenario

Michael Heck*, Sakriani Sakti, Satoshi Nakamura

Augmented Human Communication Laboratory, Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan

Abstract

In this work we make use of unsupervised linear discriminant analysis (LDA) to support acoustic unit discovery in a zero resource scenario. The idea is to automatically find a mapping of feature vectors into a subspace that is more suitable for Dirichlet process Gaussian mixture model (DPGMM) based clustering, without the need of supervision. Supervised acoustic modeling typically makes use of feature transformations such as LDA to minimize intra-class discriminability, to maximize inter-class discriminability and to extract relevant informations from high-dimensional features spanning larger contexts. The need of class labels makes it difficult to use this technique in a zero resource setting where the classes and even their amount are unknown. To overcome this issue we use a first iteration of DPGMM clustering on standard features to generate labels for the data, that serve as basis for learning a proper transformation. A second clustering operates on the transformed features. The application of unsupervised LDA demonstrably leads to better clustering results given the unsupervised data. We show that the improved input features consistently outperform our baseline input features.

Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords: acoustic unit discovery; Bayesian nonparametrics; Dirichlet process; feature transformation; Gibbs sampling; unsupervised linear discriminant analysis; zero resource

1. Introduction

In a zero resource scenario, large amounts of labeled training data, parallel data, and knowledge about the target language are unavailable for developing speech processing systems with supervised techniques. Where infants are capable of robustly modeling acoustic and language models in an unsupervised way, current speech technology is not yet capable to imitate these capacities.

Confronted with an unknown language, human experts usually attempt to define a set of acoustic units that fully covers the underlying sound repertoire. Core techniques of machine learning approaches to this task are pattern

* Corresponding author. Tel.: +81-743-72-5265 ; fax: +81-743-72-5269.
E-mail address: michael-h@is.naist.jp

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.032

matching [1,2] on raw audio data and unsupervised sound unit detection [3] and have already been successfully applied to solve tasks such as spoken term detection [4], topic segmentation [5] or document classification [6].

In non-clinical situations where development data is usually unavailable, model complexity is not known a priori. Bayesian models such as the Dirichlet process Gaussian mixture model (DPGMM) automatically adjust the model complexity given the data. DPGMMs have been successfully applied to speech processing tasks such as unsupervised lexical clustering [7]. Previous work [8] clustered standard MFCC speech features by inferring a DPGMM. Each Gaussian was interpreted as modeling a specific sound class. The posteriorgrams were evaluated to show that DPGMM is a suitable technique to automatically detect sound classes in untranscribed data.

It is straightforward to assume that more advanced feature representations may lead to a better classifier performance. For instance, context information is an important factor to correctly classify speech features in common speech processing systems. Chen et al. [8] use MFCC features with first and second derivatives for clustering. The derivatives help cover a small context but triple the dimensionality. Feature stacking can cover a much larger context, but at significantly higher expenses in terms of dimensionality. A feasible processing of high-dimensional feature vectors makes dimension reducing feature transformations mandatory.

Traditional supervised acoustic modeling typically makes use of feature transformations such as linear discriminant analysis (LDA) [9] to minimize intra-class discriminability, to maximize inter-class discriminability and to extract relevant informations from high-dimensional features spanning larger contexts. Class discriminating properties of feature vectors are critical for clustering. However, LDA needs class labels to estimate the feature transformations, making it difficult to use in a zero resource setting where the classes and even their amount are unknown.

In this work we attempt to improve the DPGMM clustering by introducing unsupervised LDA to the sampling pipeline. There has been work that utilize k-means clustering to automatically obtain pseudo labels for LDA estimation [10,11]. We similarly attempt to automatically produce class labels, but we want to overcome the limitation of having to predefine the size of the label set. For that, we use a non-parametric DPGMM sampler to generate labels for our untranscribed data. Our contribution is an easy to understand two-staged clustering framework that automatically finds a dynamically sized set of framewise class labels for unsupervised LDA transformation to project high-dimensional large context covering feature vectors into a more suitable subspace for DPGMM clustering.

2. Dirichlet Process Gaussian Mixture Model

DPGMMs (also known as infinite GMMs) extend finite mixture models by the aspect of automatic model selection: The model finds it's complexity automatically given the training data. Inference is typically sample based using a Markov chain Monte Carlo (MCMC) scheme such as Gibbs sampling. The sampler used here is combining a restricted Gibbs sampler with a split/merge sampler. For more in-depth informations, please refer to [12,8].

2.1. Generative process

Let $X = x_1, \cdots, x_n$ be a set of observations. The generative process of X given a DPGMM is as follows:

- Mixing weights $\boldsymbol{\pi} = \{\pi_1, \cdots, \pi_k\}$ are generated according to a stick-breaking process
- GMM parameters $\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_k\}$ are generated according to a prior distribution $\mathrm{NIW}(m_k, S_k, \kappa_k, \nu_k)$
- A label z_i is assigned to every x_i, according to $\boldsymbol{\pi}$
- A data point x_i is generated according to the z_i-th Gaussian component

$\theta_k = \{\mu_k, \Sigma_k\}$ are Gaussian parameters, and the parameter set of the prior Normal-inverse-Wishart (NIW) distribution consists of a prior m_0 for μ_k, a prior S_0 for Σ_k, the belief-strength κ_0 in m_0 and the belief-strength ν_0 in S_0.

2.2. Inference

The parallelizable sampler used here alternates between a non-ergodic restricted Gibbs sampler and a split/merge sampler to form an ergodic MCMC sampler.