Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 46 (2017) 233–248

www.elsevier.com/locate/csl

Scalable algorithms for unsupervised clustering of acoustic data for speech recognition[☆]

Shakti P. Rath

HLT Department, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis (South Tower), Singapore 138632

Received 25 October 2016; received in revised form 25 April 2017; accepted 2 June 2017
Available online 13 June 2017

Abstract

In this paper an unsupervised clustering algorithm is developed for acoustic data in the context of speech recognition tasks. One of the key features of the algorithm is scalability to large data sets. Specifically, given the unlabeled training and test sets, the class-labels of the utterances are obtained in an automatic manner. The extracted labels may correspond to the speakers in the speech corpus if the data is relatively clean. The proposed scheme is attractive from an industrial perspective as it alleviates the need to store the speaker-labels manually, saving considerable amount of human efforts and expenses. The core of the algorithm comprises a three-stage architecture that processes the input data one after the other, while each stage is designed to perform a well-defined and specific task. In more detail, the first-pass involves a bottom-up clustering mechanism, the second-pass comprises a cluster splitting operation and the third-pass consists of a cluster refining process. Each of the stages allows for data parallelization using multiple CPUs that leads to faster computation. Two alternative forms of the algorithm are presented – the first considers Gaussian distributions and the other i-Vectors – to facilitate the clustering. Although the algorithm may find applications in various realms of speech recognition, in this paper, the effectiveness of the schemes are evaluated by means of speaker adaptive training (SAT) and speaker-aware training of DNN-HMM acoustic models. In particular, experiments are conducted on the Switchboard task to extract the speaker-labels for the utterances in the training and test sets. It is shown that the SAT DNN-HMM trained using the Gaussian based scheme yields a 7.2% relative improvement in the ASR accuracy over the speaker independent DNN-HMM, whereas the i-Vector approach provides an additional improvement, amounting to a 10.8% relative gain overall. The standard SAT DNN-HMM developed using the ground-truth speaker-labels is found to be only 2.7% relative better than the proposed scheme. Similar observation is made as with speaker-aware training. The analysis of computational complexity, conducted stage by stage, demonstrates the scalability of the proposed algorithms.

Keywords: Unsupervised clustering; Bottom-up clustering; Large scale training; DNN-HMM; Speaker adaptive training; i-Vector

1. Introduction

In the context of automatic speech recognition (ASR), the knowledge of class-labels for the speech utterances, which indicates some form of homogeneity of the utterances, can be leveraged in several ways to improve the quality

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

E-mail addresses: shaktir@i2r.a-star.edu.sg, shaktirath@gmail.com

http://dx.doi.org/10.1016/j.csl.2017.06.001

of acoustic model. For instance, many standard speech corpora distributed by Linguistic Data Consortium[1] provide the speaker-ids, indicating the identity of the speakers enunciating the utterances. In this case speaker adaptive training (SAT) can be applied to improve the ASR accuracy over the speaker independent (SI) system (Gales, 1998; Anastasakos et al., 1996). In other cases, the environment and background noise labels can be exploited to apply noise-robust techniques (Gales, 2001; Wang and Gales, 2012; Gales and Young, 1996; Acero et al., 2000; Seltzer and Acero, 2012).

In recent years there has been significant progress in neural network based acoustic modeling. High computational power and improved machine learning techniques have made it possible to train neural network with “deep” architecture under feasible development time constraints. Moreover, training of such deep networks over very large data set has also been made possible. Over wide variety of applications many speech research groups have demonstrated the efficiency of deep neural network (DNN) (Hinton et al., 2012; Seide et al., 2011; Jaitly et al., 2012; Liao et al., 2013), and more recently recurrent neural network (RNN) (Saon et al., 2014), convolutional neural network (CNN) (Sainath et al., 2013; Sercu et al., 2016), Long Short-Term Memory (LSTM) (Graves et al., 2013; Sak et al., 2014; Hochreiter and Schmidhuber, 1997; Sainath et al., 2015), and Time Delay Neural Network (TDNN) (Peddinti et al., 2015) as acoustic modeling devices.

It is also well established that SAT applied by means of constrained maximum likelihood linear regression (CMLLR) (Gales, 1998), yields a large gain in the DNN-HMM accuracy over the speaker-independent case (Seide et al., 2011; Jaitly et al., 2012; Rath et al., 2013; Miao et al., 2014). Often, in this large scale set-ups keeping track of the speaker-labels of the audio segments while collecting the speech data might become time-consuming and expensive. The lack of this information therefore will prevent the application of SAT, limiting the performance of the ASR system to that of the SI DNN-HMM, which is undesirable. In order to be able to apply SAT in such scenario, the speaker-labels need to be computed from the utterances (segments) automatically, the process which is referred to as unsupervised clustering. The purpose of clustering is to group the utterances in the corpus into acoustically homogeneous groups.

In this paper an unsupervised clustering algorithm is developed targeting very large acoustic data sets. One of the highlights of the algorithm is scalability to large data sets. The core of the algorithm comprises of first a bottom-up clustering mechanism and secondly a cluster refining process. It works in a three-stage frame-work. In particular, steered by a hierarchical bottom-up clustering mechanism, the *first-pass* merges the utterances in the given audio into a small number of broad acoustic classes. Two different implementations are considered for the first-pass − in the first Gaussian models are used to represent the cluster distributions, whereas in the other i-Vectors are utilized to express them compactly in a low-dimensional Euclidean space. As these models may capture different intrinsic properties of the utterance groups, the composition of the clusters learned by both are likely to be very different. In either of the two cases, the clusters generated by the first-pass are supplied as the input to the second-pass. In the *second-pass*, the first-pass clusters are further divided into much larger number of granular classes. Finally, the *third-pass* aims to refine the second-pass clusters to extract more discriminative cluster definitions, while employing complex and powerful distribution models, namely, Gaussian mixture models (GMMs), to learn the cluster distributions.

The main attributes of the proposed algorithm include computational efficiency and scalability to large data sets. Each stage of the three-stage architecture is assigned to perform a specific task, while data processing in every stage is parallelized using multiple central processing units (CPUs), which, in overall, leads to a linear reduction in total processing time. Given a dataset, the number of clusters to create is controlled via various tunable parameters. These parameters may also be adjusted to strike a balance between computational complexity and class-resolution power of the algorithm. Irrespective of the quantity of data to be clustered and the number of clusters needed in the end, the algorithm works in a consistent manner. The proposed scheme presents significant industrial interests as it alleviates the need to store the speaker-labels manually, saving considerable amount of human efforts and expenses.

The performance of the scheme is validated by conducting experiments on the Switchboard corpus, which is a large vocabulary continuous speech recognition task wherein telephone conversations between speakers are recorded at a sampling rate of 8 KHz. Since the dominant source of variability in this task is the speaker factor, the class labels produced by the automatic clustering are likely to correspond to the speakers in the corpus. In separate experiments, the extracted labels are utilized to apply speaker adaptive training (SAT) to DNN-HMM acoustic model using

[1] https://www.ldc.upenn.edu