On classification and segmentation of massive audio data streams

Charu C. Aggarwal

Received: 1 December 2007 / Revised: 21 May 2008 / Accepted: 22 August 2008 / Published online: 16 October 2008
© Springer-Verlag London Limited 2008

Abstract In recent years, the proliferation of VOIP data has created a number of applications in which it is desirable to perform quick online classification and recognition of massive voice streams. Typically such applications are encountered in real time intelligence and surveillance. In many cases, the data streams can be in compressed format, and the rate of data processing can often run at the rate of Gigabits per second. All known techniques for speaker voice analysis require the use of an offline training phase in which the system is trained with known segments of speech. The state-of-the-art method for text-independent speaker recognition is known as Gaussian mixture modeling (GMM), and it requires an iterative expectation maximization procedure for training, which cannot be implemented in real time. In many real applications (such as surveillance) it is desirable to perform the recognition process in online time, so that the system can be quickly adapted to new segments of the data. In many cases, it may also be desirable to quickly create databases of training profiles for speakers of interest. In this paper, we discuss the details of such an online voice recognition system. For this purpose, we use our micro-clustering algorithms to design concise signatures of the target speakers. One of the surprising and insightful observations from our experiences with such a system is that while it was originally designed only for efficiency, we later discovered that it was also more accurate than the widely used GMM. This was because of the conciseness of the micro-cluster model, which made it less prone to over training. This is evidence of the fact that it is often possible to get the best of both worlds and do better than complex models both from an efficiency and accuracy perspective. We present experimental results illustrating the effectiveness and efficiency of the method.

Keywords Classification · Segmentation · Audio streams
1 Introduction

The problem of speaker voice analysis and classification is useful in a number of applications such as real time monitoring, detection, and surveillance. In this paper, we are concentrating on the problem of text-independent speaker classification in which the actual textual content of the speech is not available for modeling purposes. A number of statistical and machine learning methods have been recently proposed for speaker classification. Some examples of such techniques may be found in [10].

A well-known method for speaker classification and identification is that of Gaussian mixture modeling (GMM) [11,12]. The first step is to extract multi-dimensional feature vectors in order to represent portions of sampled speech. In this method, it is assumed that each data point extracted from the speech segments from a number of known speakers are used to estimate the parameters of a GMM model. Then the data points from an unknown speech segment are applied to each speaker model in order to estimate the maximum likelihood fit. The model with the highest fit is reported as the relevant class label. The GMM modeling method has been widely popular because of its intuitive appeal and effectiveness, and has therefore been used extensively. A number of other pattern recognition models which are often used for speaker classification are discussed in [4,6,7,10,14].

In many applications, it is desirable to perform the speaker identification in real time. In such cases, we need to have a system which is adaptive enough to learn characteristics of new speakers in real time, and used these learned profiles in order to construct the final model for speaker classification. Unfortunately, popularly used models such as GMM are not very appropriate for real time speaker modeling. This is because the first step of determining the parameters of the mixture model requires an iterative computationally intensive approach known as the EM algorithm. The second stage of model fitting requires the evaluation of a log likelihood criterion on the data set for each model. These techniques can be computationally intensive in practice, and are often difficult to use effectively for real time modeling and classification. Furthermore, we are also interested in other mining variations of the classification problem in which it is desirable to model or match individual segments of speech from unknown speakers in real time. This is not possible with an iterative approach such as the EM algorithm. In addition, since we are making the stream assumption for all data processing, we assume that each point in the data can be scanned only once throughout the computation. This assumption is also violated by all other speaker identification systems [4,6,10] known to us. Therefore, we need to construct a system which can work with the constraints of a one-pass system and still provide accurate results for speaker identification.

In addition to speaker classification, we would like to design a system which is capable of detecting quick changes in the pattern of the underlying data stream. Such a system is useful in applications in which it is desirable to segment portions of speech into different speakers. We note that this segmentation may need to be done in an unsupervised way, since example segments of the different speakers may not be known in advance. In order to design a system which can work with such challenges, we construct a number of techniques which are designed towards stream based online processing of massive amounts of audio data. We adapt our earlier research results for stream micro-clustering [2] in order to create fast cluster based signatures of the data. These signatures are constructed and used in real time in order to perform the speaker identification.

We also report a number of interesting observations about the behavior of the micro-clustering model. While our original aim was only to design an efficient online system for speaker recognition, the results also turn out to be qualitatively superior to the GMM model. This is because the GMM model requires a significantly large number of training parameters.