Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 48 (2006) 85–95
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Gaussian-selection-based non-optimal search for speaker identification

Marie Roch *

San Diego State University, 5500 Campanile Drive, San Diego, CA 92182-7720, USA

Received 3 February 2004; received in revised form 28 March 2005; accepted 5 June 2005

Abstract

Most speaker identification systems train individual models for each speaker. This is done as individual models often yield better performance and they permit easier adaptation and enrollment. When classifying a speech token, the token is scored against each model and the maximum a priori decision rule is used to decide the classification label. Consequently, the cost of classification grows linearly for each token as the population size grows. When considering that the number of tokens to classify is also likely to grow linearly with the population, the total work load increases exponentially.

This paper presents a preclassifier which generates an *N*-best hypothesis using a novel application of Gaussian selection, and a transformation of the traditional tail test statistic which lets the implementer specify the tail region in terms of probability. The system is trained using parameters of individual speaker models and does not require the original feature vectors, even when enrolling new speakers or adapting existing ones. As the correct class label need only be in the *N*-best hypothesis set, it is possible to prune more Gaussians than in a traditional Gaussian selection application. The *N*-best hypothesis set is then evaluated using individual speaker models, resulting in an overall reduction of workload.

Keywords: Speaker recognition; Text-independent speaker identification; Talker recognition; Gaussian selection; Non-optimal search

1. Introduction

Traditionally, speaker identification is implemented by training a single model for each speaker in the set of individuals to be identified. Identification is accomplished by scoring a test utterance against each model and using a decision rule, such as the maximum a posteriori (MAP) decision rule where the class of the highest scoring model is selected as the class label.

The advantages of such schemes over training a single model with multiple class outputs is that enrollment of new speakers does not require the

* Tel.: +1 619 594 5830; fax: +1 619 594 6746.
E-mail address: marie.roch@ieee.org

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.06.003

training data for the existing speakers and the training of individual models is faster. The downside to this approach is that the computational complexity required for identification grows linearly for the classification of each speaker as the number of the speakers increases. Unless the majority of enrolled speakers are infrequent users of the system, it is reasonable to expect that as the number of registered users increase, the number of classification requests could increase at least linearly. Under this assumption, the system load rises exponentially as the registered population increases due to the increase of requests and increase of models to check against.

In this work, we develop a preclassifier which produces a hypothesis that a token is most likely to belong to one of a small subset of the possible classes. The token is then rescored against the traditional per class models and a final class label is assigned. The design goals are to produce a preclassifier with the following properties:

(1) The computational workload of the system is reduced.
(2) Enrollment of new speakers should be of low cost. Regeneration of the preclassifier as new speakers are enrolled should not be cost prohibitive with respect to both time and space requirements.
(3) There should be minimal or small impact on the classification error rate. Use of the system should incur no more than a small penalty in classification performance.

We will show that a system meeting the above criteria can be achieved through a novel application of Gaussian selection. A Gaussian selection system is constructed which evaluates a subset of the non-outlier Gaussians near each speaker. Unlike traditional Gaussian selection systems like those described in Picheny's (1999) review of large vocabulary dictation systems, there is no attempt to capture all of the distributions for which the point is not an outlier. A small subset of the distributions is sufficient to identify a set of promising candidates. This candidate set is then rescored using speaker-specific models and the MAP decision rule is applied.

It should be noted that the proposed system is non-optimal; there is no guarantee that the hypothesis set will contain the correct class of the token being classified. As will be demonstrated empirically in the experimental section, in most cases this has minimal impact on the classification error rate for the evaluated corpora.

The remainder of this paper is organized as follows: Section 2 describes an overview of existing techniques to reduce the computational load of speaker recognition systems. Next, we review Gaussian selection (Section 3) and present a way to specify the outlier test in terms of probability. Section 4 describes how Gaussian selection can be applied to construct an efficient preclassifier which meets the stated design goals. Section 5 describes the experimental methodology, and results are reported in Section 6. Finally, we summarize our findings in Section 7.

2. Background

The proposed Gaussian-selection-based preclassifier differs from other non-optimal search techniques such as the well-known beam search (Huang et al., 2001), time-reordered beam search of Pellom and Hansen (1998), and confidence-based pruning (Kinnunen et al., in press) in that classification speed can be increased without pruning candidates before all feature vectors are considered, but the system could of course be combined with such techniques. In terms of system architecture, the proposed technique is similar to the two stage classification system of Pan et al. (2000) where two learning vector quantizers (LVQ) of differing size are used for first and second pass scoring.

There are several partition-based approaches that have been proposed for speech and speaker recognition systems. The partitioning results in a pruning of Gaussian evaluations. These systems can be separated into techniques which provide separate partitions for each model and those that partition the entire feature space. Pruning of Gaussians in these techniques occurs when computing the posterior likelihood for the final classification decision, and in all cases it is not part of a N-best match strategy.