Detecting Goats in Speaker Verification Systems

Orith Toledo-Ronen and Hagai Aronowitz

IBM Research – Haifa, Haifa University Mount Carmel, Haifa 31905, Israel
{oritht, hagaia}@il.ibm.com

Abstract

We present a method for detecting goats in a text-dependent speaker verification system using only the enrollment data. The goat detection process is based on extracting an appropriate feature from each enrollment session and ranking all the enrollment sessions in the system according to this feature. The lowest-ranking sessions, which are likely to have a high false-reject rate, are selected. We present results of the goat detector on a text-dependent speaker verification system.

1. Introduction

In the context of biometric systems, a goat is a person that is hard to recognize and is often denied access to the system. Since Doddington et al. [1] first introduced the biometric menagerie in 1998, some work has been done on detecting goats. Several studies, including Doddington’s paper, have shown that a small number of speakers in a speaker verification system account for a disproportional amount of errors. This phenomenon is common to different types of biometric modalities including voice, fingerprint, iris, and face recognition.

Approaches for detecting goats are based on statistical tests of score variability and using ranking techniques to identify the poorly-performing speakers [2][3] or to classify the speakers into the different menagerie classes [4]. Other methods are focused on user-specific score normalization or user-specific decision strategies to reduce the effect of individual speakers on the system and by that to improve the overall system’s performance [5][6].

In an operational speaker verification system, it is important to automatically detect the poor performers and to take action accordingly. If a user is suspected of being a goat, we would like to identify this as early as possible, preferably after the enrollment session and before verification. Upon detection of a suspected problematic user, the user may be handled by repeating the enrollment process, by verification using a different modality, or by using a more sophisticated algorithm. The overall false rejection rate could then be improved and, in particular, customer satisfaction levels could be maintained for goat users.

In this work, we attempt to automatically detect the users that have difficulties in being verified, in particular identifying speakers that have a high false reject rate based on the enrollment data of the speaker. Ideally we would like to detect the goats during the enrollment session. Alternatively, the system should be able to detect goats within the first few verification attempts. In this work we focus on goat detection during enrollment with limited amount of data of the enrollment session.

The rest of the paper is organized as follows. In Section 2 we describe the method for detecting the poorly performing users. In Section 3 we present the experimental setup and results, and in Section 4 we provide the conclusions of this work.

2. Method

Our approach for goat detection is by ranking all the available enrollment sessions of all the speakers in the system and selecting the subset of poorly performing sessions. The process consists of three steps. First, a feature is extracted from each enrollment session. Then all the enrollment sessions are ranked based on this feature and 1% of the sessions with the lowest ranks are selected. This subset represents the sessions suspected to have poor performance. We then evaluate the goat detector by measuring the performance on the data in the selected subset. With this method, we detect unsuccessful enrollments. Given the enrollment session of a speaker, we are trying to predict failure in future verification attempts of the same speaker, or in other words, we attempt to detect the goats with respect to a particular enrollment session. The enrollment feature we use for goat detection is based on variations of the leave-one-out method described in [7]. Each enrollment session in our data set consists of 3 repetitions of the password. The feature is called Mean11, and is extracted from the enrollment session by the following algorithm:
1. Enroll with one repetition from the enrollment session and verify on the other two repetitions.
2. Repeat Step 1 with the other two repetitions from the enrollment session.
3. Compute the average of the unnormalized verification scores from Step 1 and 2.

In case that the verification scores are symmetric, the average score for Mean11 is computed on 3 verification scores.

3. Results

3.1. Dataset description

The data set we use for authentication is text-dependent and consists of a common text for both enrollment and verification. The data was collected by the Wells Fargo (WF) Bank. The WF corpus is based on audio
recordings of WF employees, and it consists of several common texts from 750 speakers. The data is divided into a development set (200 speakers) and an evaluation set (550 speakers). Each speaker in the corpus recorded two sessions in landline phone and two sessions in cellular phone. Each recording session consists of 3 repetitions of the password. In this work we report results on one password, a 10-digits string of counting from 0 to 9. Our evaluation set consists of 527 speakers with a total of 2,035 recorded sessions. Overall we have 17,994 target trials and 193,884 impostor trials for evaluation. The evaluation trials cover the mismatched channel condition, and the impostor trials have the same gender as the target.

3.2. Experimental Setup

Our speaker verification system is a GMM-supervector-based system, which is based on a UBM-GMM system similar to the system described in [8]. We use a front-end with Mel-frequency cepstral coefficients (MFCCs) extracted every 10ms with a 25ms window. An energy-based voice activity detector is applied to remove non-speech frames. The final feature set consists of 13 cepstral coefficients augmented by 13 delta cepstral coefficients. Feature warping is applied with a 300 frame window before computing the delta features. GMMs with 512 gaussians are mean-adapted from a UBM. The UBM is trained from the same text as used during enrollment. We compensate intra-speaker inter-session variability using a variant of the nuisance attribute projection (NAP) method called two-wire NAP [9]. The compensated supervectors are scored using the C_GM inner-product proposed in [10], and the scores are normalized by ZT-norm.

3.3. Experimental Results

We evaluate the proposed goat detection algorithm by comparing the performance of the evaluation trials corresponding to the subset of the 1% lowest-ranking enrollment sessions that are automatically selected by the goat detector with the overall performance of all the evaluation trials of all the enrollment sessions. Figure 1 shows the DET curves of the 1% lowest-ranking sessions and the 1% highest-ranking sessions selected by the Mean11 enrollment feature described in Section 2. We see that the miss probability or the false reject rate of the selected subset of suspected goats is higher than the overall system false reject rate. For example, at the EER point the false reject of the selected 1% lowest-ranking sessions is 7.8% compared to the overall rate of 2%. In addition, we see that the automatically selected 1% highest-ranking enrollment sessions have a low false reject rate. In [11], we explored several other enrollment features and we show comparative results.

4. Conclusions

We presented a method for detecting goats in a text-dependent speaker verification system based on quality feature of the enrollment session itself, without the need for additional verification data from the true speaker. We presented results on the enrollment-based feature exploiting information from the target speaker enrollment data.

5. Acknowledgements

The authors wish to thank Wells Fargo for collecting and providing the data for this study.

6. References