Towards Goat Detection in Text-Dependent Speaker Verification

Orith Toledo-Ronen[1], Hagai Aronowitz[1], Ron Hoory[1], Jason Pelecanos[2], David Nahamoo[2]

[1] IBM Research – Haifa, Haifa University Mount Carmel, Haifa 31905, Israel
[2] IBM T.J. Watson Research Center, Yorktown Heights, NY, U.S.A
{oritht, hagaia, hoory}@il.ibm.com, {jwpeleca, nahamoo}@us.ibm.com

Abstract

We present a method that identifies speakers that are likely to have a high false-reject rate in a text-dependent speaker verification system ("goats"). The method normally uses only the enrollment data to perform this task. We begin with extracting an appropriate feature from each enrollment session. We then rank all the enrollment sessions in the system based on this feature. The lowest-ranking sessions are likely to have a high false-reject rate. We explore several features and show that the 1% lowest-ranking enrollments have a false reject rate of up to 7.8%, compared to our system's overall rate of 2.0%. Furthermore, when using a single additional verification score from the true speaker for ranking, the false-reject of the 1% lowest-ranking sessions rises up to 33%.

Index Terms: speaker verification, text dependent, failure to enroll, goat detection

1. Introduction

In voice biometric systems, the performance varies for different users. A well known paper on this topic by Doddington et al. [1] describes the four different characteristics of speakers and calls it the "Biometric menagerie". A *sheep* is the default speaker type that dominates the population and has good performance, while a *goat* is a speaker that is hard to recognize and has a disproportional share of the false reject errors. The other two speaker types are related to the false accept rate. A *lamb* is a speaker that is easy to imitate and a *wolf* is a speaker that is successful at imitating others. Doddington's paper shows that only a small number of enrolled speakers have a relatively higher rate of recognition errors compared to the others. They found that 25% of the most goat-like speakers in their speaker verification dataset accounted for 75% of the false reject errors. Several other studies 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. An extension of the biometric menagerie is described in a recent paper [2].

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.

An early work on goat detection is described in [3] for a VQ-based speaker identification system. Speakers are ranked based on a speaker specificity measure combining intra-speaker and inter-speaker variability, and some correlation between the specificity measure and the recognition error rate is shown. However, the intra-speaker variability measure is based on test utterances from the speaker.

Another method for model quality evaluation during enrollment is described in [4]. It removes non-representative utterances from the enrollment session and replaces them by new utterances. However, it requires a large amount of enrollment data from each user, which we don't have in real applications, and it aims at improving the quality of the enrollment model while we are focusing on detecting the goats given the enrollment model. In [5], an interesting approach for classifying the speakers into the different menagerie classes is presented. The classification is based on ranking the scores of the speakers in a speaker identification paradigm. Although ranking the scores is similar to our method, this paper proposed a method for classifying the speakers based on the evaluation scores while we rank the scores from the enrollment session for predicting or detecting potentially problematic users. Koolwaaij et al. [6] presented a related work on evaluating the quality of the model for speaker verification. Their technique uses the mean LLR score of the target training utterances after Z-norm as an indicator of the discriminative ability of the model. The paper mentions that with a threshold on this quality score from enrollment, a decision can be made on whether re-enrollment of the model is needed, but without showing its ability to detect goats. The paper focuses on using this quality measure for weighting the scores from individual models for improving the recognition performance.

Another related work on goat detection by Poh and Kittler is described in [7]. In that work, the methodology of ranking the speakers and evaluating the goats is similar to our method. However, the ranking method is applied for face, iris, and fingerprint biometrics while we apply it for voice biometrics. Moreover, in one part of our work we propose ranking features that are based entirely on the enrollment data and do not require additional data from the user as done in that related work. In [8], the authors are looking at similarity between speakers.

Other methods are taking a different approach of user-specific score normalization or user-specific decision strategies to reduce the effect of individual speakers on the system and thus to improve the overall system's performance. For example, F-norm [9][10] uses additional client data for estimating the target score distribution and simultaneously aligning the target and impostor score distributions.

In this work, we attempt to automatically detect the goats. In other words, we aim to find users that have difficulties in being verified, in particular identifying speakers that have a high false reject rate based on very little data from the speaker. Ideally we would like to detect the goats during the enrollment session. Alternatively, the system should detect goats within the first few verification attempts. This is a challenging task given the small amount of data available for making the decision. Other factors, such as channel mismatch, affect the performance and make the goat detection even more difficult, although some of these factors are not directly related to the speaker's characteristic as a goat.

In this paper we use a ranking-based method for identifying the low-performing speakers. Our method is based on ranking the enrollment sessions of all the speakers in the system, selecting a subset of the lowest-ranking sessions, and measuring the performance on the selected subset. We apply this technique to text-dependent speaker verification. In such applications the enrollment session is usually very short and only few repetitions (typically 3) of the vocal password are available for deciding on the *quality* of the enrollment session and an even harder problem is whether information from a single session can predict the performance of the speaker in future verification attempts.

The rest of the paper is organized as follows. In Section 2 we describe the method for measuring the enrollment quality and 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 and selecting the subset of poorly performing sessions. The rank of each enrollment session is based on features that are extracted from the enrollment session itself and provide an indication of the *quality* of the enrollment. We extract several different features from the enrollment session and measure their performance in separating the goats from the sheep. For each feature, we measure the performance on the selected subset of enrollment sessions and compare it to the overall performance. The process consists of three steps. First, a feature is extracted from the 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 performance on the data of the selected subset. During the evaluation part, we distinguish between the failure-to-enroll and the goat detection scenarios, as described in detail in 2.2. In both cases, the ranking process is identical, but they differ in their evaluation methodology.

2.1. Enrollment Features

Each enrollment session in our data set consists of 3 repetitions of the password. We investigated several enrollment features, some of which are based on variations of the leave-one-out method described in [11].

- **Mean21**: enroll with 2 repetitions from the enrollment session and verify on the 3rd repetition; a total of 3 scores per enrollment; ranking is based on the average of these scores. The scores are not normalized.
- **Mean11**: enroll with 1 repetition from the enrollment session, verify on the other 2 repetitions; a total of 3 (symmetric) scores per enrollment; ranking is based on the average of these scores. The scores are not normalized.
- **Z-norm**: estimate the Z-norm parameters for each enrollment session by testing the enrollment model on trials from the development set and use the Z-norm distribution mean for ranking.
- **F-ratio**: a combination of Mean11 and Z-norm based on the measure described in [9].

2.2. Evaluation methods

During evaluation, we apply two different methods. The first method is *failure-to-enroll* detection, which means detecting unsuccessful enrollments. This term is most commonly related to low quality inputs, describing the cases that features cannot be extracted from a particular enrollment session. Some simple quality checks can be applied and as a result a recommendation for re-enrollment could be provided to the user. In this paper we focus on a different aspect of enrollment failure, which is predicting failure in future verification attempts for a given enrollment session or in other words, the detection of a goat with respect to a particular enrollment session. The second evaluation method is *goat detection*. In this evaluation a goat is detected if the performance of trials formed from all the enrollment sessions of the same speaker (excluding trials from the session used for ranking) show low performance.

Our baseline evaluation consists of verification trials from all combinations of each speaker's enrollment and test sessions. At each verification trial one session is used for enrollment and the test is performed on data from another session. This means that if each speaker has 4 recorded sessions, we will have 4 enrollments for that speaker and we will test each enrollment session on data from the other 3 test sessions. Given that an enrollment session *E* is selected by the ranking, we define the following two methods for evaluating the quality of the ranking:

1. **Failure-to-enroll**: evaluating all the verification trials performed on the selected enrollment session *E*.
2. **Goat detection**: evaluating all the verification trials from sessions of the same speaker as in the selected enrollment session *E*, excluding verification trials on data from the selected enrollment session *E* itself.

The first method represents the simple scenario of detecting bad enrollment sessions by testing only on the selected enrollment sessions. The second method evaluates the goat detection in a broader sense, in which, when a particular enrollment session of a user is selected, we also measure its effect on the speaker's performance when enrolling and testing with other (independent) sessions. This way we can examine our technique for detecting the poorly performing speakers from all of their data and not only in relation to a specific enrollment session.

Table 1 explains the evaluation schemes for failure-to-enroll detection (FTE) and for goat detection (GD). In the example shown in the table, a speaker has 3 sessions: E1, E2, and E3. In FTE, the evaluation is done only on tests performed on the selected enrollment session and in the GD evaluation, testing is done on all the sessions of the same speaker, excluding the selected enrollment itself.

Enrollment	Test FTE	Test GD
E1	E1-T2, E1-T3	E1-T2, E1-T3, E2-T3, E3-T2
E2	E2-T1, E2-T3	E2-T1, E2-T3, E1-T3, E3-T1

Table 1: Example of failure-to-enroll and goat detection evaluation schemes (E1-T2 means enrollment on session 1 and testing with session 2).

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, which are divided into a development set (200 speakers) and an evaluation set (550 speakers). Each speaker recorded 2 landline phone sessions and 2 cellular phone sessions, and each session consists of 3 repetitions of the password. In this work we report results on one password, a 10 digits string (counting from 0 to 9). Our evaluation set consists of 527 speakers with 2,035 sessions. It contains 17,994 target trials, 193,884 impostor trials, and a total of 211,878 trials. The evaluation trials cover the mismatched channel condition, and the impostor trials have the same gender as the target. This data set was used for all our experiments with the enrollment features described in Section 2.1. In addition, we examined the case that a score from a single verification attempt of the true speaker for each enrollment session is available, and we use it for ranking. The verification score in this case is obtained not from the enrollment session but from a repetition of the password from another session. Therefore, for evaluating the test score ranking, we set aside one session from each speaker and used it for ranking purposes. This, of course, limited the amount of data we had for evaluation in this experiment because the ranking session was not part of the evaluation. For the other experiments, with features that are extracted directly from the enrollment session itself, we used all the evaluation data available for each speaker without dedicating one session for ranking.

3.2. Experimental Setup

We use a GMM-supervector-based system which is based on a UBM-GMM system similar to what is described in [12]. The front-end is based on Mel-frequency cepstral coefficients (MFCCs). An energy-based voice activity detector is used to locate and remove non-speech frames. The final feature set consists of 13 cepstral coefficients augmented by 13 delta cepstral coefficients extracted every 10ms using a 25ms window. Feature warping is applied with a 300 frame window before computing the delta-features. GMMs of order 512 are adapted from a UBM with mean adaptation only. The UBM is trained from the same text as used in enrollment.

We compensate intra-speaker inter-session variability using a variant of the nuisance attribute projection (NAP) [13] method called 2-wire NAP [14], which removes not only an intra-speaker variability subspace but also a subset spanned by the eigenvectors corresponding to the largest eigenvalues of the inter-speaker covariance matrix. The compensated supervectors are scored using the C_{GM} inner-product proposed in [15]. Finally we normalize the scores using ZT-norm.

3.3. Results with Enrollment Features

Table 2 summarizes the performance of the enrollment features described in Section 2.1 on the subset of the 1% lowest-ranking sessions for the FTE and GD conditions, as described in Section 2.2. The results presented are the equal error rates (EER) on the evaluation set for the selected subset of the enrollment sessions. The false reject rates (FR), and the false accept rates (FA) at the threshold of the EER point of the overall system are also shown. The FR rate of the selected subset increases to 7.8% in the FTE evaluation and to 4.6% in the GD evaluation using the Mean11 feature that is based on information extracted only from the enrollment session itself. The ranking with the Z-norm feature helps to increase the FA rate, especially in the FTE evaluation. With the F-ratio feature, which combines information from both the enrollment and from the Z-norm impostor scores distribution, the FR rate goes up to 6.5% in both evaluation conditions. Figure 1 and 2 show the DET curves of the 1% lowest-ranking sessions selected by three different enrollment features for the FTE and GD evaluations. We see that the Z-norm score is stronger on the impostor side while the Mean11 is stronger on the target side and the F-norm is in between.

Evaluation Method	**Feature**	**EER (%)**	**FR (%)**	**FA (%)**
All	None	2.0	2.0	2.0
FTE	Mean21	2.6	5.0	1.6
	Mean11	3.7	7.8	1.6
	Z-norm	4.0	4.1	4.0
	F-ratio	3.6	6.5	2.3
GD	Mean21	2.4	3.5	1.7
	Mean11	2.8	4.6	1.8
	Z-norm	3.9	4.2	1.9
	F-ratio	4.2	6.5	1.8

Table 2: Performance of the enrollment features when selecting the 1% lowest-ranking enrollment sessions.

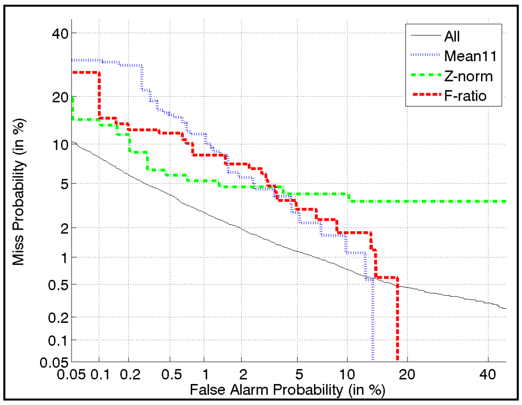

Figure 1: DET curve performance of the failure-to-enroll evaluation for three enrollment features.

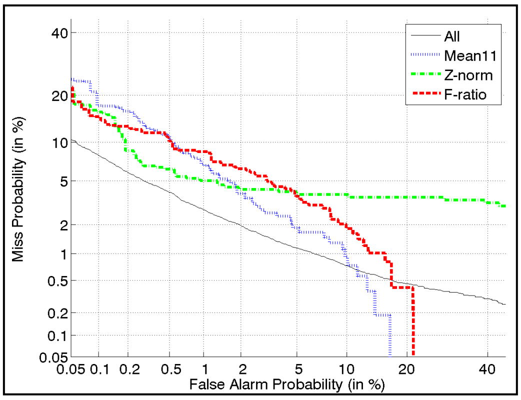

Figure 2: DET curve performance of the goat detection evaluation for three enrollment features.

3.4. Ranking with a Verification Score

If an additional utterance is available from the true speaker, we can use it in several ways. In our experiment, the additional sample is coming from a different session and not from the enrollment. The simplest approach would be to use the verification score from the additional sample for ranking with the existing enrollment model. We call this ranking "R1-test". This means that we use the score from testing R1 against the enrollment model which is trained from three repetitions (E3). However, we can also take the additional sample from the speaker add it to the model and re-enroll (E3+R1). Once we re-enroll with the four repetitions of the password, we can now apply the enrollment-based ranking on the new extended model. We tried the Mean31 feature, which is enrollment with 3 repetitions and testing the 4th, producing a total of 4 scores per enrollment session and taking the average for ranking. The Mean31 feature on the extended enrollment is analogous to the Mean21 feature on the original enrollment session. The results are summarized in Table 3 for the original enrollment (E3) and the extended enrollment (E3+R1). In both cases the evaluation set is the same, meaning excluding R1 from testing. First, the baseline is different for the two systems, and we can see the effect of additional enrollment data in improving the overall performance. Secondly we found that using the additional verification score for ranking (R1-test) on the original enrollment model (E3) is more powerful than using it or an enrollment ranking score (Mean31) on the re-enrolled model (E3+R1). In addition, we see that combining scores from different sessions, as done in the case of the Mean31 feature on enrollment E3+R1, is weaker than the Mean21 on enrollment E3 in the FTE evaluation (for making a decision with respect to a particular enrollment session) while it is stronger in the GD evaluation (for a decision with respect to several session of the same speaker).

Evaluation Method	Enroll	Feature	EER (%)	FR (%)	FA (%)
All	E3	None	2.0	2.0	2.0
	E3+R1	None	1.3	1.3	1.3
FTE	E3	R1-test	16.5	25.8	2.2
	E3	Mean21	2.1	4.2	1.7
	E3+R1	R1-test	5.0	8.3	2.0
	E3+R1	Mean31	1.2	1.7	0.7
GD	E3	R1-test	21.2	33.3	2.2
	E3	Mean21	2.4	4.0	1.7
	E3+R1	R1-test	9.8	18.8	1.5
	E3+R1	Mean31	3.3	5.6	1.3

Table 3: Performance of ranking with one verification score or with an enrollment score versus re-enrolling with the additional data and ranking with Mean31 enrollment score.

3.4.1. Validating the Results

Our results are based on experiments with a limited amount of data from each speaker. In order to validate the results, we ran an experiment of randomly selecting 1% of the enrollment sessions and evaluating the performance on the selected subset. We repeated this experiment 10,000 times and in each trial we computed the EER of the evaluation set of the selected sessions, as well as the FR and FA rates at the overall system's threshold. We found that for the FTE evaluation 0.81% of the random trials resulted in FR rate greater than the FR of the Mean11 feature. For the GD evaluation, 7.4% of the random trials were measured. As for the F-ratio feature, about 2% of the random trials had FR rate higher than the FR of the F-ratio feature in both evaluation methods.

4. Conclusions

We presented a method for detecting goats in a text-dependent speaker verification system based on quality features of the enrollment session itself, without the need for additional verification data from the true speaker. We also investigated several possibilities of using an additional sample from the true speaker either for ranking or for re-enrollment. We presented several enrollment-based quality features exploiting information from the target speaker enrollment data and information from impostor trials on a development set.

5. Acknowledgements

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

6. References

[1] G. Doddington et al., "SHEEP, GOATS, LAMBS and WOLVES A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation", in Proc. *ICSLP*, 1998.

[2] N. Yager and T. Dunstone, "The Biometric Menagerie", in *IEEE Transaction on PAMI*, Vol. 32, No. 2, 2010.

[3] J. Thompson and J. Mason, "The Pre-Detection of Error-Prone Class Members at the Enrollment Stage of Speaker Recognition Systems", *ESCA Workshop* 1994.

[4] J. R. Saeta and J. Hernando, "Model Quality Evaluation during Enrollment for Speaker Verification", in Proc. *ICSLP*, 2004.

[5] J. Koolwaaij and L. Boves, "A New Procedure For Classifying Speakers In Speaker Verification Systems", in Proc. *Eurospeech* 1997.

[6] J. Koolwaaij et al., "On model quality and evaluation in speaker verification", in Proc. *ICASSP* 2000.

[7] N. Poh and J. Kittler, "A Methodology for Separating Sheep from Goats for Controlled Enrollment and Multimodal Fusion", in Proc. of *Biometrics Symposium (BSYM)*, 2008.

[8] L. Stoll and G. Doddington, "Hunting for Wolves in Speaker Recognition", in Proc. *Odyssey* 2010.

[9] N. Poh and S. Bengio, "F-ratio Client-Dependent Normalization for Biometric Authentication Tasks", in Proc. *ICASSP*, 2005.

[10] N. Poh et al., "Group-specific Score Normalization for Biometric Systems", in *IEEE Computer Society Workshop on Biometrics (CVPR)*, 2010.

[11] Gu, Y. et al., "Speaker Verification in Operational Environments-Monitoring for Improved Service Operation", in Proc. *ICSLP* 2000.

[12] W. M. Campbell, D. E. Sturim, D. A. Reynolds, "Support Vector Machines using GMM Supervectors for Speaker Verification", in *Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[13] W. Campbell, D. Sturim, D. Reynolds, A. Solomonoff, "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation", in Proc. *ICASSP*, 2006.

[14] Y. A. Solewicz and H. Aronowitz, "Two-Wire Nuisance Attribute Projection", in Proc. *Interspeech* 2009.

[15] W. M. Campbell and Z. N. Karam, "Simple and Efficient Speaker Comparison using Approximate KL Divergence", in Proc. *Interspeech*, 2010.