SHEEP, GOATS, LAMBS and WOLVES
A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation

George Doddington[1,2,3,5]*, Walter Liggett*[1]*, Alvin Martin*[1]*, Mark Przybocki*[1]*, Douglas Reynolds*[3,4]*,*

[1] National Institute of Standards and Technology, [2] The Johns Hopkins University
[3] U.S. Department of Defense, [4] MIT Lincoln Laboratory, [5] SRI International

ABSTRACT

Performance variability in speech and speaker recognition systems can be attributed to many factors. One major factor, which is often acknowledged but seldom analyzed, is inherent differences in the recognizability of different speakers. In speaker recognition systems such differences are characterized by the use of animal names for different types of speakers, including *sheep*, *goats*, *lambs* and *wolves*, depending on their behavior with respect to automatic recognition systems. In this paper we propose statistical tests for the existence of these animals and apply these tests to hunt for such animals using results from the 1998 NIST speaker recognition evaluation.

1. INTRODUCTION

Research lore in speech and speaker recognition has for many years acknowledged the existence of striking performance inhomogeneities among speakers within a population. The terms *sheep* and *goats* have been used to characterize speakers for whom systems perform well and poorly, respectively. Little systematic study has been made up to this time, however, to characterize such differences within a population of speakers. One recent review, however, does discuss speaker performance differences, and applies animal names to problem speakers[1].

Experiments in the recognition of speech and speakers are strongly influenced by results for the most poorly performing speakers. This nonuniform performance often is an important issue in applications. Thus, in addition to characterizing general population performance in terms of miss and false alarm error rates, it is also important to characterize system robustness over the population. In a study using the 1997 NIST speaker recognition evaluation data, various different random selections of speaker populations showed a factor of 9 change in false alarm rate at a fixed miss rate[2]. Clearly, the mean population performance is not giving the complete picture.

In this study we compute and analyze population statistics for speaker recognition performance based on the test data that was used for the NIST 1998 speaker recognition evaluation. This evaluation includes data from more than 500 speakers and recognition results from 12 systems.

2. THE ANIMALS

In addition to the traditional *sheep* and *goat* populations, we can expand our hypothetical menagerie of speakers for the speaker verification task. Speaker verification is a detection task, for which system performance may be characterized in terms of two types of errors, namely misses (in which the true speaker is not detected) and false alarms (in which an impostor speaker is falsely detected). We define our menagerie as follows:

Sheep – Sheep comprise our default speaker type. In our model, sheep dominate the population and systems perform nominally well for them.

Goats – Goats, in our model, are those speakers who are particularly difficult to recognize. Goats tend to adversely affect the performance of systems by accounting for a disproportionate share of the missed detections. The goat population can be an especially important problem for entry control systems, where it is important that *all* users be reliably accepted.

Lambs – Lambs, in our model, are those speakers who are particularly easy to imitate. That is, a randomly chosen speaker is exceptionally likely to be accepted as a lamb. Lambs tend to adversely affect the performance of systems by accounting for a disproportionate share of the false alarms. This represents a potential system weakness, if lambs can be identified, either through trial and error or through correlation with other directly observable characteristics.

Wolves – Wolves, in our model, are those speakers who are particularly successful at imitating other speakers. That is, their speech is exceptionally likely to be accepted as that of another speaker. Wolves tend to adversely affect the performance of systems by accounting for a disproportionate share of the false alarms. This represents a potential system weakness, if wolves can be identified and recruited to defeat systems.

3. DISTRIBUTIONS AND TESTS

The speaker verification task is a detection task to determine whether some specified (target) speaker spoke some given segment of speech. To avoid semantic confusion, we will refer to the actual (true) speaker of the speech segment as the ***segment speaker***, and the hypothetical (target) speaker as the ***model speaker***. The speaker verification system evaluates a speaker hypothesis by scoring the given speech segment against the model for the hypothesized speaker. The system then makes a decision, based upon the resulting score: If the score is greater than some fixed threshold (which is independent of model), then the model speaker hypothesis is accepted. Otherwise the hypothesis is rejected[3].

A speaker recognition system is tested by presenting it with many segments from many (segment) speakers. Each of these segments is evaluated, both for the segment (true) speaker hypothesis and for other model (impostor) speakers. Thus the data available for our analysis are scores from a large number of trials, $\{S(i,j,k)\}$, where: S = the system output score for a trial; i = the segment index for segment speaker j; j = the segment speaker index; k = the model speaker index. For each segment speaker j, we can think of a population of speech segments, each with a corresponding score against the model k. Thus, we can think of a score probability density function for a segment speaker and model speaker, $f_s(\bullet \mid j,k)$.

From these scores, we wish to determine if there are speaker effects that demonstrate the existence of goats, lambs and wolves. In order to do that, we assert the null hypothesis namely that there are no speaker differences, and then determine whether our experimental results violate this null hypothesis. Here are the relevant distributions and null hypotheses for our menagerie:

Goats - Determine if the density of system output scores is a function of the segment speaker when the segment speaker is the model speaker. The density of interest is $f_s(\bullet \mid k,k)$ and the null hypothesis is that this density does not depend on k.

Lambs - Determine if the density of system output scores is a function of the model speaker when the segment speaker is not the model speaker. The null hypothesis is that $f_s(\bullet \mid j,k)$ does not depend on k for all j as long as $j \neq k$.

Wolves - Determine if the density of system output scores is a function of the segment speaker when the model speaker is not the segment speaker. The null hypothesis is that $f_s(\bullet \mid j,k)$ does not depend on j for all k as long as $j \neq k$.

Here are the statistical graphical analysis tests that we used:

Goats - First, using scores for which j = k, S(i,j,j), compute variances from sets of scores attributable to the same segment speaker and check to see if these variances depend on j. Second, compute means from the same sets of scores, and check to see if these means depend on j. We do this by comparing the means with 2.5 and 97.5 percentiles under the assumption that the means and the variances do not depend on j.

Lambs - For model k, plot maximum score obtained as

$$\max_{\{i,j \mid j \neq k\}} S(i,j,k)$$

versus each corresponding score for which j = k, S(i,k,k).

Wolves - Compute maximum scores obtained as

$$\max_{\{k \mid k \neq j\}} S(i,j,k),$$

and use them as in the goat case. First, using the maximum scores, compute variances from sets of maximum scores attributable to the same speaker and check to see if these variances depend on j. Second, compute means from the same sets of scores, and check to see if these means depend on j. We do this by comparing the means with 2.5 and 97.5 percentiles under the assumption that the means and the variances do not depend on j.

F-test: This is the standard one-way analysis of variance statistical test of whether there is a population (in this case speaker) effect. We applied it to test for potential goats using all segment scores for each speaker. In testing for lambs and wolves, all the scores of the segments corresponding to a particular segment j and model speaker k were first averaged to give $\bar{S}(j,k)$. The test for lambs then used the sample $\{\bar{S}(j,k) : \text{all } j \neq k\}$ for each model speaker k, while the test for wolves used the sample $\{\bar{S}(j,k) : \text{all } k \neq j\}$ for each model speaker j.

Kruskal-Wallis Test: This is a non-parametric one-way analysis of variance by ranks test[4] [section 6.2]. For goats, all same speaker scores are used (limited to speakers with at least five test segments), while for lambs and wolves, the multiple segment scores for each segment and model speaker pair are averaged as above before applying the test. The test assigns ranks to all of the averaged scores under consideration, and the ranks for each sample (corresponding to a hypothesized speaker of a particular species) are summed. Use of a non-parametric test avoids an assumption of normality in the data, which is system dependent.

Durbin Test: This is a variant of the Friedman two-way analysis of variance by ranks test[4] [section 7.1], modified to allow for an incomplete block design[4] [section 7.4]. This is appropriate for considering scores where the segment and model speakers are different (testing for existence of lambs and wolves) and the data may be viewed as conditioned on the two different types of speaker. Averaging across segments was performed as above. The Durbin test assigns ranks to the averaged scores of each segment speaker (lamb test) and each model speaker (wolf test). These rankings are summed for each model speaker (lamb test) and each segment speaker (wolf test).

4. DATA

The hunt for the animals was conducted using data segments from the 1998 NIST Speaker Recognition Evaluation[3,5]. The evaluation speech data is derived from the Switchboard-II, phase 2 corpus and consists of 500 speakers (250 male, 250 female), three training conditions, three test utterance length conditions and 5000 tests per condition.

To eliminate many of the confounding variables which are known to cause performance differences among sub-populations but are not directly attributable to speaker differences, such as handset mismatches, we restricted our analysis to a subset of the entire evaluation. Specifically, we used results from the female speakers, for models built from two-session training, for 30-second test segments from different phone numbers than for training. Also, both test and training segments were limited to data (automatically) determined to be from electret microphones. This data set consisted of 535 trials with matching segment and model speakers from 154 speakers, and 4763 trials with non-matching speakers, involving 221 segment speakers.

Results are available from the 12 participating sites. However, for the official evaluation, only 10 model speakers were scored against each test segment. This limits the number of trials to be used for the lamb and wolf testing. To expand the analysis data, the MIT Lincoln Laboratory system[6] was run again to produce scores for all test segments against all model speakers. This increased the number of trials with non-matching speakers to 399,462 and the number of speakers to 240. It is this system's results that form the basis of the analysis, although we expect the analysis to be generally true since most participating sites used a system similar to the MIT LL system.

5. ANALYSIS

Goats Analysis - For speakers with at least two test segments, the distribution of the variances of the same speaker model scores was found to be consistent with the assumption of equal variance for each speaker. Figure 1 shows the distribution of mean scores for the speakers, plotted with the number of test segments of the speaker on the horizontal axis. Were there no dependence on speaker, only one point in twenty would lie outside the 2.5 and 97.5 percentiles shown. The speakers below the 2.5 percentile can reasonably be considered goats.

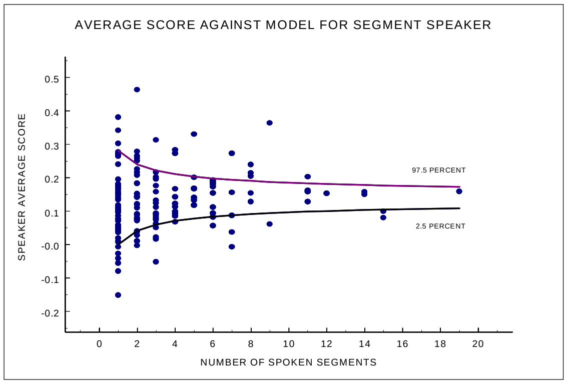

Figure 1: Average true speaker scores with 2-sided 5% critical values.

The Kruskal-Wallis test was applied to the 39 speakers with at least five test segments as the segment speaker.[7] [p. 89]. Both the F-test (analysis of variance) and the Kruskal-Wallis test applied to these samples easily yielded rejection of the null hypothesis at the 0.01 significance level. Thus the existence of goats in the speaker population, at least for the system used to generate the scores, is affirmed.

Lamb Analysis - Figure 2 shows for each model, the score for each segment by the model speaker and the largest score for segments by imposters. One model gave a very high score for an imposter segment. Otherwise, there is no evidence of a separate sub-population of models that could be considered lambs. The models with large maximum imposter scores do exhibit lamb-like behavior.

There are 221 female speakers with models from electret training data. For each of these, and for each female speaker of electret segments (of which there are 241), we found the mean of the scores involving the given segment speaker and the given model. We then considered the set of mean scores for each of the 221 model speakers. Both the F-test (analysis of variance) and the Kruskal-Wallis test easily found significance at 0.01 significance level, supporting the conclusions that the 221 samples of means scores could not be regarded as coming from a common distribution. Thus the existence of lambs is supported in this sense.

There are 219 female speakers with both electret models and electret test segments. From these we generated a square matrix of mean scores with the diagonal entries omitted. This is an incomplete block design to which we may apply the Durbin test, essentially a Friedman two–way analysis of variance by ranks test. Unsurprisingly, this also readily shows significance at the 0.01 significance level.

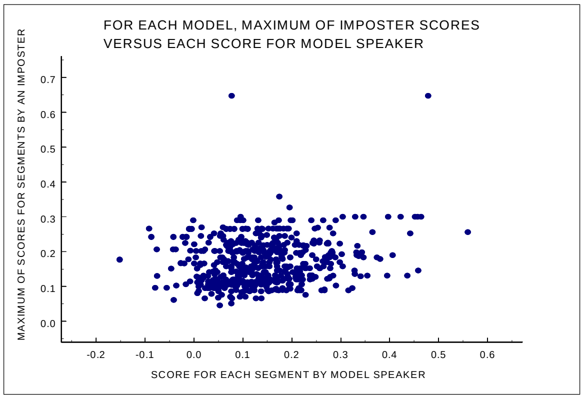

Figure 2: Scatterplot of models: Highest model score vs. segment model score.

Wolf Analysis - Figure 3 is analogous to Figure 1 with substitution for the Figure 1 scores, the maximum over models for which the segment speaker does not match the model speaker. With this substitution, we still have instances of values (maximum scores) by the same speaker. Thus, we can and did check that the data are consistent with the assumption of equal variance for each speaker. Figure 3 shows the mean maximum scores for speakers plotted with the number of test segments of the speaker on the horizontal axis. Were there no dependence on speaker, only one point in twenty would lie outside the 2.5 and 97.5 percentiles shown. The speakers above the 97.5 percentile can reasonably be considered wolves. In addition, applying the Durbin test to the matrix of scores from the 219 speakers with electret train and test data also rejected the null hypothesis at a 0.01 significance level.

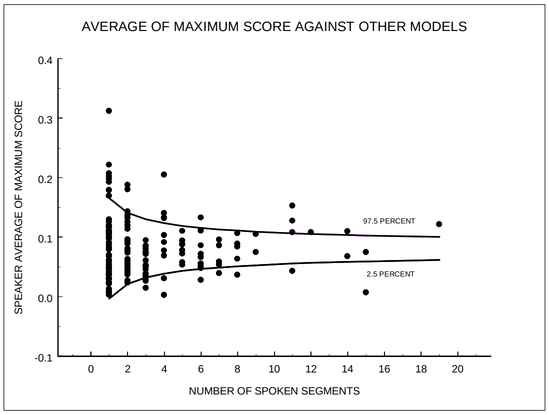

Figure 3: Maximum scores with 2-sided 5% critical values

Correlation among animals - The Durbin test assigns a rank sum to each speaker corresponding to her scores as a model speaker (lamb test) and a rank sum corresponding to her scores as a segment speaker (wolf test). Since both rank sums are based on scores where the segment and target speakers are different, it is perhaps reasonable that they should be mildly correlated (correlation coefficient ~ 0.26). For the 39 female speakers considered in the tests for goats, there appears to be no correlation between the goat with that of the lambs or wolves rankings provided by the statistical tests.

Correlation among systems - For the tests where the segment and model speakers were identical (possible goats) we also have complete results from all of the automatic systems that participated in the 1998 NIST evaluation. Figure 4, shows the normalized rank sum for the 39 female speakers included in the Kruskal-Wallis test for five of these systems. The speakers are ordered by their normalized rank sum for system 1, which is similar to the system used for the other results in this paper.

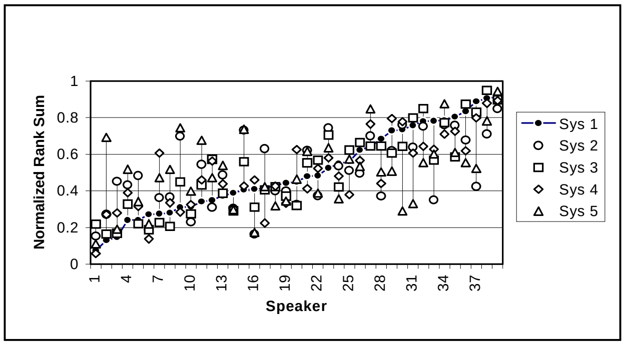

Figure 4: Kruskal-Wallis Normalized Rank Sums for 39 speakers and five systems.

6. AN EXAMPLE OF ERROR COUNTS

Using the statistical tests to label speakers as goats, lambs and wolves, we examined their contribution to speaker verification errors. We ranked the 39 speakers used in the Kruskal-Wallis test according to how *goat-like* they were. From the wolf and lamb test, we ranked the 219 speakers used in the Durbin test according to how *wolf-like* and *lamb-like* they were.

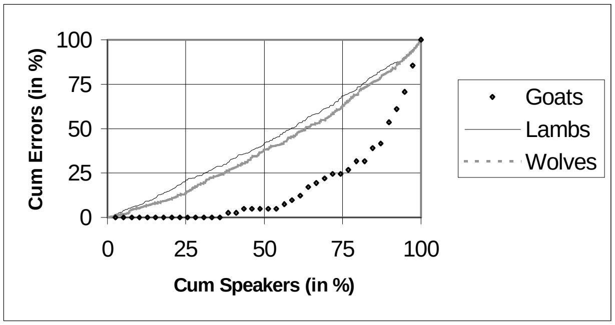

Figure 5: Cumulative error distributions for rank-ordered speakers.

At a posterior operating point of Pr[false alarm] = 10% and Pr[miss] = 1%, we then analyzed the errors attributed to each animal type. Figure 5 shows the cumulative error distribution of the rank ordered speakers. It appears that the goats have the greatest performance effect, with 25% most goat-like speakers contributing 75% of the miss errors.

7. CONCLUSIONS

In this paper we have considered three aspects of speaker differences on the performance of a speaker recognition system. We label the speakers contributing to these effects goats, lambs and wolves. We have tested whether these effects are real, and we have found that they are. Note, however, that simply rejecting the hypothesis that there is no effect does not prove the existence of distinct classes of speakers. In fact, this seems quite unlikely. More likely is that the population of speakers exhibits a continuum of goatish, lambish and wolfish characteristics. It is also quite possible that the speaker differences that we have found are a result of dependencies that are not directly attributable to the speaker, *per se*. For example, there may be casual dependencies between speaker identity and the type of phone used. Nonetheless, considering that we have demonstrated significant speaker differences, it remains to develop standard meaningful characterizations of these differences, and to include in future evaluations measures of system robustness to these differences.

8. REFERENCES

[1] J.P. Campbell, “Speaker Recognition: A Tutorial”, Proc. of the IEEE, vol. 85, no. 9, Sept. 1997

[2] MIT LL Site Presentation, 1997 NIST Speaker Recognition Workshop, June 1997.

[3] NIST 1998 Speaker Recognition Evaluation Plan, http://jaguar.ncsl.nist.gov/evaluations/speaker/feb98/plans/current_plan.htm

[4] W.W. Daniel, Applied Nonparametric Statistics, Houghton Mifflin Company, 1978

[5] M.A. Przybocki and A. F. Martin, "NIST Speaker Recognition Evaluations", Proceedings, LREC, Granada, Spain, 1998, 331-335

[6] D.A. Reynolds, “Comparison of Background Normalization for Text-Independent Speaker Verification,” Eurospeech, 1997

[7] G.K. Kanji, 100 Statistical Tests, SAGE Publications, 1993