The NIST Speaker Recognition Evaluations: 1996-2001

Alvin F Martin, Mark A. Przybocki

National Institute of Standards and Technology
Gaithersburg, MD 20899 USA
alvin.martin@nist.gov mark.przybocki@nist.gov

Abstract

We discuss the history and purposes of the NIST evaluations of speaker recognition performance. We cover the sites that have participated, the performance measures used, and the formats used to report results. We consider the extent to which there has been measurable progress over the years. In particular, we examine apparent performance improvements seen in the 2001 evaluation. Information for prospective participants is included.

1. Introduction

NIST (The National Institute of Standards and Technology) has coordinated evaluations of text independent speaker recognition using conversational telephone speech over the past six years. Some discussion of these evaluations may be found in [1], [2], and [3]. These evaluations have had as primary objectives:

- Exploring promising new ideas in speaker recognition,
- Developing advanced technology incorporating these ideas, and
- Measuring the performance of this technology.

Key features of these evaluations have been that they be:

- Simple,
- Focused on core technology issues,
- Fully supported, and
- Accessible.

The evaluations have all included the basic one-speaker detection task consisting of a series of trials. Each trial presents the system with a target speaker, defined by some speech by the speaker (usually two minutes in duration), and with a test segment of up to one minute in duration, spoken by a single unknown speaker. For each trial, the system must decide whether or not the unknown speaker is the target, producing both a yes-or-no hard decision and a likelihood score.

There are two types of trials: *target trials* where the unknown speaker is the target, and *non-target trials* where the unknown speaker is someone else. System errors for the first type are *misses*; for the second type *false alarms*. System performance may then be characterized by the two error rate types: *miss rate* and *false alarm rate*. The requirement for likelihood scores for all trials using a common scale allows these two error rates to be determined at multiple system operating points. [8]

The 1999-2001 evaluations have included additional tasks beyond that of one-speaker detection. These tasks have been set in the context of test segments containing speech by multiple speakers. See [4] for further information. This paper restricts its discussion to the one-speaker detection task included in all of these evaluations.

Most of the data used in these evaluations have come from the Switchboard Corpora of conversational telephone speech, available from the Linguistic Data Consortium (LDC) [5]. The 2000 and 2001 evaluations also used, in a separate test, non-conversational telephone speech from the Castilian Spanish AHUMADA Corpus [6] made available by Javier Ortega-Garcia of the Universidad Politecnica de Madrid.

2. Evaluation Participants

As shown in Table 1, the participants over the past six years have been from 12 countries on five continents, making these truly worldwide evaluations.

In some cases two or more participants have worked in cooperation while submitting individual results for separate systems. Most notable has been the ELISA Consortium. This is an organization of primarily European sites that has created common system components while allowing individual sites to pursue variants to some components of particular interest to them.

3. Evaluation History

The present basic format of these evaluations using conversational telephone data was adopted in 1996. Subsequent evaluations have increased the numbers of speakers and trials, and have added other tasks as mentioned above. The AHUMADA data was added in 2000, and some of the newly collected Switchboard cellular data was included in 2001.

Each evaluation has included certain types of trial specified as being the primary condition of particular interest. For example, earlier evaluations included segments of three different durations, namely 3, 10, or 30 seconds, and those of one particular duration were specified to be part of the primary condition. Likewise, there were multiple types of training data for each target speaker, with one specified as primary.

From the beginning, it was recognized that different telephone handsets could greatly affect recognition performance. In particular, target trials would be easier if the training and test handsets used by the target speaker were identical. Both same and different target trial handsets were part of the primary condition in different years.

It also became apparent, largely because of work done by MIT-Lincoln Lab, a participating site, that the microphone type of the handset was an important factor in performance. In the United States both electret and carbon button microphones are common. Performance is affected both by type (electret microphones enhance performance) and by whether the training and test types are the same. More recent evaluations have specified electret type as part of the primary condition.

Table 2 lists the primary conditions, numbers of speakers and trials, and some of the distinguishing features of the several evaluations. It should be noted that the relatively large numbers of speakers and trials have been distinguishing features of the NIST evaluations. This has enhanced confidence not only in the overall evaluation results, but in examinations of various contributing factors that involve observing performance on small subsets of the data. These subsets need to be large enough for meaningful results, as suggested by Doddington's "Rule of 30" (see [7]):

> *To be 90 percent confident that the true error rate is within +/- 30 percent of the observed error rate, there must be at least 30 errors.*

4. Presentation of Results

The official performance measure for the NIST evaluations has been a weighted average, denoted C_{DET}, of the miss and false alarm error rates as defined in figure 1.

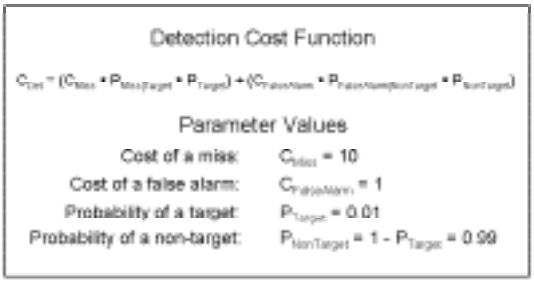

Figure 1: C_{DET} function with current parameter values.

The primary means of presenting system performance, however, has been with the use of DET (Detection Error Tradeoff) Curves [8]. These show the full range of possible operating points of the system based on the likelihood scores each system provides for all trials. A typical plot of such curves is shown in figure 2. The use of a normal deviate scale on both axes results in the curves being (approximately) linear if the underlying distributions of likelihood scores for both the target and non-target trials are (approximately) normal.

Note that specific operating points may high-lighted on each DET Curve. Generally the point that correspond to the actual (hard) decisions (denoted by '•') and the point on the curve for which the C_{DET} value is minimal (denoted by '♦') are plotted. A good choice of likelihood threshold value for the actual decisions will result in these points being identical.

The C_{DET} values corresponding to these points are then sometimes shown in a bar chart plot as in figure 3.

5. Measuring Progress

A key question of interest is whether, and how much, progress in recognition performance has been achieved over the course of the evaluations. This can be frustratingly difficult to determine accurately. Although all evaluations included the basic task of one speaker detection, there have been both major and minor changes from one evaluation to another in the primary recognition condition of interest, for which sites were asked to optimize their systems (see table 1). Moreover, different test sets, even though selected in exactly the same manner, can easily be quite different in inherent difficulty.

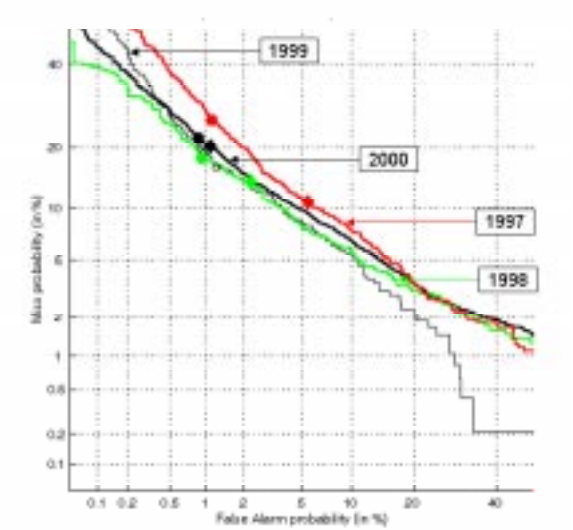

Figure 2: A typical DET Curve plot is shown. As noted below, these are actually performance results for one site in successive years.

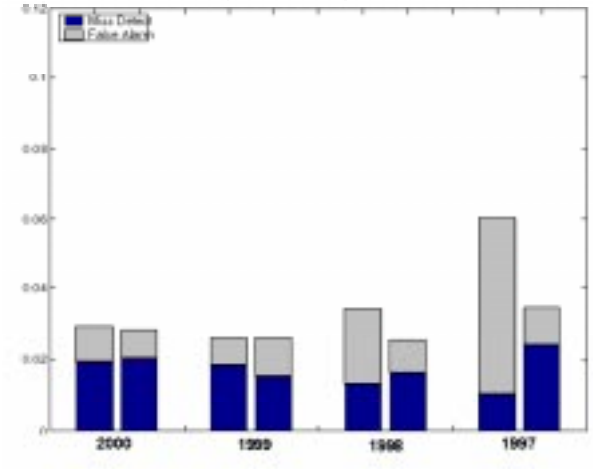

Figure 3: A typical bar graph plot corresponding to the DET Curves of figure 2. The left bar in each pair shows hard decision cost; the right bar the minimum C_{DET}. The lower part of each bar shows the cost of missed detections; the upper part the cost of false alarms.

This has been noted in recent NIST coordinated evaluations of speech recognition [9]. Figure 2 in fact shows performance results for one site from 1997-2000. For each year, the plot shown is for the subset of the 1997 and 1998 subsets were basically the same, there were unavoidable changes in segment durations and training procedures in 1999 and 2000, confounding performance comparisons. From figure 3 it is clear that the site did improve its threshold setting procedure over this period, producing actual C_{DET} values better approximating the minimum values.

The best indicator of performance improvement can be observed when a site provides results for both a previous and a current system on a given test set. This has been available in limited instances. Figure 4, for example shows performance on the 1999 primary condition data for systems from one site (different from the figure 2 site) developed for the 1997, 1998,

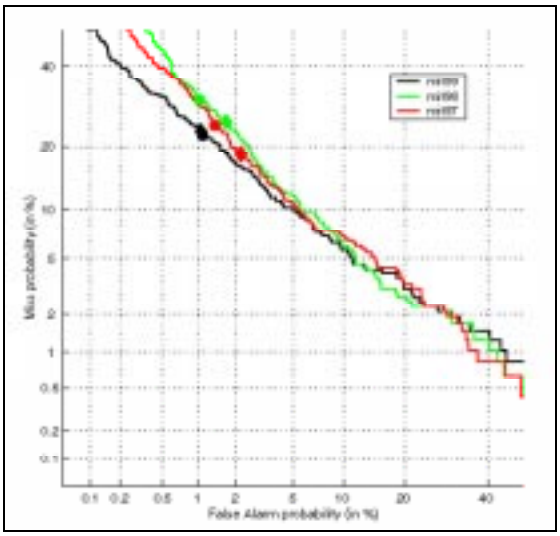

Figure 4: One site's progress from 1997-1999.

and 1999 evaluations. The three DET Curves show evidence of real, if small, performance improvement from 1997 to 1999.

The 2001 main one-speaker detection test set was primarily a repetition of that for 2000. Some additional trials using the same speakers and test segments were also included. A small additional test set involving newly collected cellular phone data was added as well.

The decision to rely primarily on a repeat of the 2000 test was made because of the lack on large quantities of fresh test data. This is a continuing problem for ongoing evaluations. But it did offer the advantage of identical test sets with which to measure year-to-year system progress. On the other hand, there is a legitimate concern that systems may have adapted themselves to the old data. It is generally believed that the large size of the test set limits the extent to which this is likely to be the case, but this requires further examination in the future.

Figure 5 shows DET Curves for 2000 and 2001 on the same set of trials for systems of six sites. Subject to the caveat noted above, there certainly appears to have been significant improvement by each of these sites.

The 2001 evaluation also included what was known as the extended data task, which used the original Switchboard-1 Corpus. Here the test segments consisted of single entire conversation sides, with speaker training data consisting of 1 to 16 such entire conversation sides of the given speaker. Moreover, participants could use machine generated transcripts of all of these conversation sides as part of their systems. Dragon Systems provided transcripts created by their ASR system for this purpose.

The extended data task was included as a result of some work by George Doddington and others showing the possibility that dramatic progress on the speaker detection task might be obtained by using such extended data including transcripts. The initial evaluation results appear very promising [10], [11]. This could represent a significant performance breakthrough for those limited applications for which such extended data would be available.

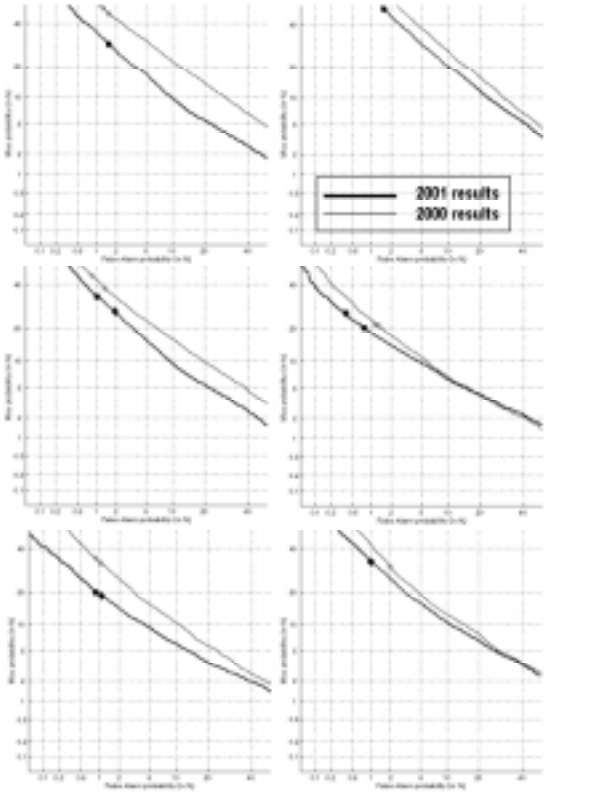

Figure 5: Comparative performance in 2000 and 2001 of systems from six sites on the identical set of trials.

6. Future Plans

NIST is now considering plans for the tests to be included in the 2002 evaluation. Suggestions in this regard are welcome.

Especially welcome would be leads and suggestions on appropriate conversational telephone type data that might be available for use.

The NIST evaluations are open to all, and new participants are welcome. Potential participants may obtain data sets from previous evaluations for development work. These data sets are available from the LDC or from NIST. Sites that are not LDC members are asked to sign a license agreement limiting data use to research purposes over a specified time period.

Evaluation information is available on the NIST web site:

http://www.nist.gov/speech/tests/spk/index.htm

References

[1] Przybocki, M. and Martin, A., NIST speaker recognition evaluation – 1997", *RLA2C*, Avignon, April 1998, pp. 120-123

[2] Doddington, G., et al., "The NIST speaker recognition evaluation – Overview, methodology, systems, results, perspective", *Speech Communication* 31 (2000), pp. 225-254

[3] Martin, A. and Przybocki, M., "The NIST 1999 Speaker Recognition Evaluation - An Overview", *Digital Signal Processing*, Vol. 10, Num. 1-3. January/April/July 2000, pp. 1-18

[4] Martin, A. and Przybocki, M., "Speaker Recognition in a Multi-Speaker Environment", *Proc. Eurospeech '01*

[5] Switchboard Corpora are available from the LDC at *http://www.ldc.upenn.edu/*

[6] J. Ortega-Garcia et al., "AHUMADA: A Large Speech Corpus in Spanish for Speaker Identification and Verification", *Proc. ICASSP '98*, Vol. II, pp. 773-776

[7] Doddington, G., "Speaker recognition evaluation methodology: a review and perspective", *RLA2C*, Avignon, April 1998, pp. 60-66

[8] Martin, A., et al., "The DET curve in assessment of detection task performance". *Proc. EuroSpeeech* Vol. 4 (1997), pp. 1895-1898

[9] Fiscus, J., et al., "2000 NIST Evaluation of Conversational Speech Recognition Over the Telephone", *Proc. 2000 Speech Transcription Workshop, http://www.nist.gov/speech/publications/tw00/html/cts10/cts10.htm*

[10] Doddington, G., "Some experiments on Idiolectal Differences among Speakers", http://*www.nist.gov/speech/tests/spk/ 2000/doc/n-gram_experiments-v06.pdf*

[11] G. Doddington, "Speaker Recognition based on Idiolectal Differences between Speakers", *Proc. Eurospeech '01*

Africa	Datavoice (South Africa)	
Asia	Amdocs (Israel)	Indian Institute of Technology Madras
Australia	Defence Science & Technology Organization	Queensland University of Technology
Europe	Ensigma (UK) Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (France) ELISA Consortium: Ecole Polytechnique Federale de Lausanne (Switzerland) Ecole Nationale Superiere des Telecommuniciations (France) Faculté Polytechnique de Mons & Rice University (Belgium and USA) Institut Dalle Molle d'Intelligence Artificielle Perceptive (Switzerland) Insitut de Recherche et Informatique et Systemes Aleatoires (France) Laboratoire Informatique d'Avignon (France) Brno University of Technology (Czech Republic)	Laboratoire Informatique de Paris (France) Nijmegan University (Netherlands)
North America	Atlantic Coast Technologies (USA) Aegir Systems (USA) Dragon Systems (USA) INRS-Telecom (Canada) Nuance (USA) RGMM (Department of Defense, USA) SRI International (USA)	Air Force Research Laboratory (USA) BBN Technologies (USA) ITT (USA) MIT-Lincoln Laboratory (USA) Oregon Graduate Institute (USA) Rutgers University (USA)

Table 1: Participating sites in the NIST Speaker Recognition Evaluations, 1996-2000.

Year	Primary Condition	Target Speakers/ Target Trials	Evaluation Features
1996	Not defined	40/3999	Tests of 3 durations, 3 training conditions Switchboard-1 data
1997	Train/test using different handset 30 second durations Train on 1 min. from each of two conversations using different handsets	~400/3050	Tests of 3 durations, 3 training conditions Switchboard-2 Phase 1 data
1998	Train/test using same handset 30 second durations Train on 1 min. from each of two conversations using same handset	~500/2687	Tests of 3 durations, 3 training conditions Switchboard-2 Phase 2 data Handset type detector info made available
1999	Train/test use different electret handsets Test durations 15-45 seconds Train on 1 min. from each of two conversations using same handset	233/479	Added multi-speaker tasks Variable durations used in main test trials Switchboard-2 Phase 3 data
2000	Train/test use different electret handsets Test durations 15-45 seconds Train on 2 min. from one conversation	804/4209	Resegmented 1997, 1998 test data for reuse Extra test on AHUMADA Spanish data
2001	Train/test use different electret handsets Test durations 15-45 seconds Train on 2 min. from one conversation	804/4209	Repeated 2000 main test with some additional trials Additional test on Switchboard cellular data Additional test allowing human or machine transcripts with extended training data

Table 2: Information on the evaluations, 1996-2001.