NIST Speaker Recognition Evaluation Chronicles

Mark Przybocki, Alvin Martin

Speech Group, Information Access Division, Information Technology Laboratory
National Institute of Standards and Technology, USA
mark.przybocki@nist.gov, alvin.martin@nist.gov

Abstract

NIST has coordinated annual evaluations of text-independent speaker recognition since 1996. During the course of this series of evaluations there have been notable milestones related to the development of the evaluation paradigm and the performance achievements of state-of-the-art systems. We document here the variants of the speaker detection task that have been included in the evaluations and the history of the best performance results for this task. Finally, we discuss the data collection and protocols for the 2004 evaluation and beyond.

1. Introduction

The Speech Group at the National Institute of Standards and Technology (NIST) has been coordinating yearly evaluations of text-independent speaker recognition technology since 1996 [1]. The evaluations have been posed primarily as detection tasks relying on various conversational telephone speech corpora (*see section 3*) as the main source of evaluation data.

During the nine years of NIST Speaker Recognition evaluations the tasks have evolved and in some cases they have run their course (i.e., tracking from 1999-2001, and segmentation from 2000-2002). But the basic task of speaker detection (determining whether or not a given speaker is speaking) has remained the primary focus of all the NIST Speaker Recognition Evaluations.

By providing explicit evaluation plans, common test sets, standard measurements of error, and a forum for participants to openly discuss algorithm success and failures (see [2]), the NIST series of Speaker Recognition Evaluations (SRE's) has provided a means for recording the progress of text-independent speaker recognition performance.

2. Evaluation Measures

Test trials for the speaker detection task can be categorized as either target trials (the specified speaker is speaking in the test segment) or impostor trials (the specified speaker is not speaking in the test segment). Each trial requires two outputs from the system under test. These are an *actual decision*, which declares whether or not the test segment contains the specified speaker, and a *likelihood score*, which represents the system's degree of confidence in its actual decision. This can result in two types of actual decision errors, *missed detections* and *false alarms*. The *miss rate* ($P_{Miss|Target}$) is the percentage of target trials decided incorrectly. The *false alarm rate* ($P_{FA|Impostor}$) is the percentage of impostor trials decided incorrectly.

2.1. C_{Det} Cost Function

NIST uses a cost function as the basic performance measure. The *C_{Det} cost* is a weighted sum of the two error rates. It is defined as the cost of a missed detection error multiplied by the miss rate, multiplied by the assumed a priori probability of a target trial, plus the cost of a false alarm error multiplied by the false alarm rate, multiplied by the a priori probability of an impostor trial (see equation 1).

$$C_{Det} = (C_{Miss} * P_{Miss|Targ} * P_{Targ}) + (C_{FA} * P_{FA|Impostor} * P_{Impostor})$$

Equation 1: C_{Det} cost function

For the NIST evaluations the cost of a missed detection has been set as 10 and the cost of a false alarm as 1. The a priori probability of a target trial has been assigned the value 0.01. Note that this does not reflect the actual target richness of the evaluation data trials.

This cost function is made more intuitive by normalizing it so that a system with no discriminative capability would have a cost of 1.0. Since deciding "false" for every trial results in an unnormalized cost of 0.1, while deciding "true" for every trial results in an unnormalized cost of 0.99, we normalize the C_{Det} values by a factor of 0.1.

2.2. Equal Error Rate

An alternative performance measure for detection tasks is the equal error rate. This is the miss (and false alarm) rate at the operating point where the two error rates are equal.

Although this is a very intuitive measure, the NIST evaluations have chosen to focus attention around a different operating point which may be more appropriate for certain applications.

2.3. DET Curves

In addition to the single number measures of C_{Det} cost and equal error rate, more information can be shown in a graph plotting all the operating points of a system. An individual operating point corresponds to a likelihood threshold for separating actual decisions of true or false. By sweeping over all possible threshold values all possible system operating points are generated.

NIST has used a variant of the popular receiver operating characteristic (ROC) curve, suggested by Swets [3], where the two error rates are plotted on the ***x*** and ***y*** axes on a normal deviate scale. NIST introduced the use of such Decision Error Tradeoff (DET) Curves [4] in the 1996 evaluation [5], and DET Curves have since been widely used for the representation of detection task performance.

Since the C_{Det} value and equal error rate correspond to points on the DET Curve, they can be marked with special symbols for easy identification.

3. Corpora for NIST SRE's

Without data there would be no research. There would certainly not be any form of evaluation. NIST has benefited from the ongoing collections of conversational telephone speech by the Linguistic Data Consortium [6]. The several collections of Switchboard style corpora, each of which included hundreds of speakers and thousands of conversations, have been extensively used in the detection tasks of the NIST Speaker Recognition Evaluations. Table 1 lists the corpora used in each of these evaluations.

Year	Corpus	Detection Tasks	Unique Attributes
1996	SWBD I	1sp lim	USA coverage
1997	SWBD II phase 1	1sp lim	Mid-Atlantic
1998	SWBD II phase 2	1sp lim	Mid-West
1999	SWBD II phase 3	1sp lim	South
2000	*recycled p1 & p2*	1sp lim	---
2000	AHUMADA	1sp lim	Spanish
2001	*repeat of 2000*	1sp lim	---
2001	SWBD I	1sp ext	---
2002	SWBD cellular p1	1sp lim	Cellular GSM
2002	SWBD p2 & p3	1sp ext	---
2002	FBI Voice DB	1sp mm	Multi Modal
2003	SWBD cellular p2	1sp lim	Cellular CDMA
2003	*repeat of 2002*	1sp ext	---
2004	MIXER	1sp var	Multi-language/ transmission types

Table 1: Corpora used for various NIST Speaker Recognition evaluations. Abbreviations: "lim" for limited-data, "ext" for extended-data, "var" for limited and extended combined, "mm" for multi-modal, and "p" for phase.

4. Evaluation Tasks

This section reviews the performance history for several variants of the speaker detection task included in the NIST evaluations over the eight year period 1996-2003.

4.1. One-Speaker Detection with Limited Data

This is the basic version of the task that has been a part of all the NIST Speaker Recognition evaluations. As implied by the name, this task evaluates a system's ability to determine if a specified target speaker is the speaker speaking in a given test segment when both the training data and test segment data contain speech from a single speaker Furthermore, "limited data" implies that the quantity of both model training and test data is restricted. This has meant in practice no more than about two minutes of speech for training, and no more than one minute for test segments.

4.1.1. Training Data

In general, approximately two-minutes of training speech data has been provided to create each target speaker model. The composition of these two minutes was varied in the early evaluations in order to analyze its effects on speaker detection performance. It was taken from a single or from two different conversations, and the two conversations were chosen to come from the same or from different telephone handsets.

From 2000 to 2003, the training data consisted two minutes of speech data from a single conversation.

4.1.2. Test Data

The first several evaluations used fixed duration test segments of either 3, 10 or 30 seconds of speech. Beginning in 1999, variable duration segments up to a maximum of one minute, with an average of about 30 seconds, were used.

4.1.3. History – State-of-the-Art

The performance history of one-speaker detection with limited data was published previously in 2002 [7]. In figure 1 we update this history of state-of-the-art performance with results from the 2003 evaluation.

It has become the accepted community practice not to publicize winners and losers as such by identifying participating sites with their performance results in open meetings and publications. This is intended to encourage evaluation participation by various sites, perhaps using high-risk techniques, without the concern of public embarrassment. As part of its agreement to participate in the NIST Speaker Recognition evaluations each site agrees that it is free to publicly present its own results, but that it may not directly compare its results to those of the other participants. In figure 1 we plot a single DET curve representing the best performance achieved in each individual year's one-speaker detection with limited data task. We try to plot evaluation conditions that are most similar across the different years. Note that better performance corresponds to a shift towards the lower left corner. Solid lines represent evaluations that used landline data, while broken lines represent evaluations that used cellular data. The actual decision operating points are shown as triangles.

This history plot represents an effort to indicate the extent of progress achieved over time. This is always problematic, however, as the task conditions as well as the specific data change with each evaluation. For the landline evaluations, there appears to be real progress shown from 1996 to 1997, and from 1997 to later years. The apparently strong 1998 results are misleading, however, as this was the only landline evaluation that used target trials in the primary testing condition where the training and test segment telephone handsets were the same.

There is greater comparability among the cellular data evaluations of 2001-2003, where protocols remained similar. In particular, the 2003 evaluation essentially reused the data from 2002, though new segments were selected from the same conversations. This choice was forced by the non-availability of new data, but as a result, the performance curve differences in figure 1 do correspond basically to system algorithm

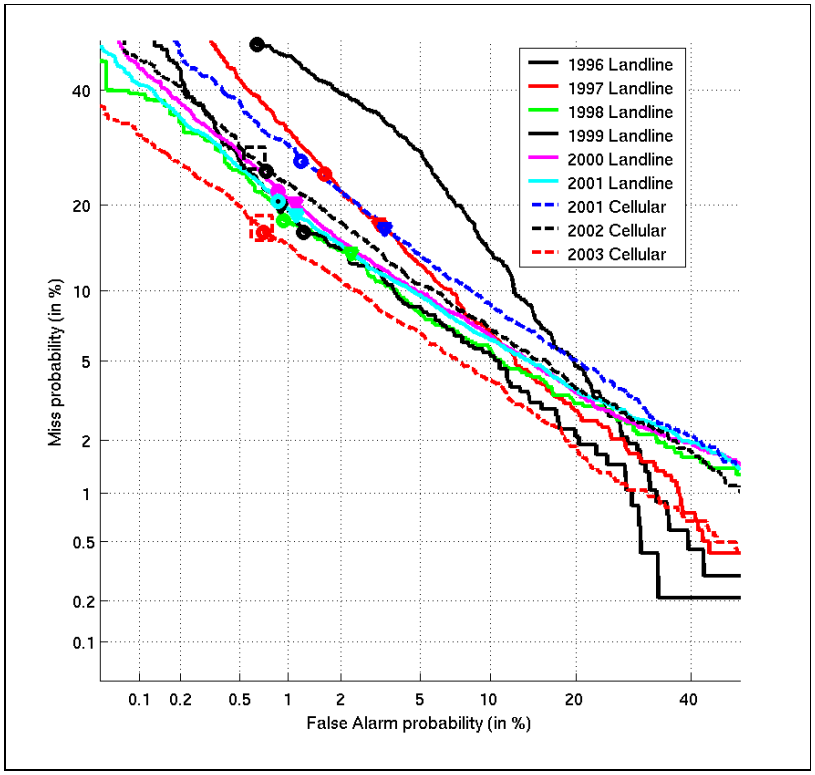

Figure 1: Best system performance history for one-speaker detection with limited data task

changes for these two years. For these two curves figure 1 includes "confidence" boxes around the actual decision points. These show the 95% confidence intervals about the miss and false rates, assuming that the target and non-target trials each represent bernouilli processes, thus giving some indication of the significance of the performance change.

It may be noted that the best comparisons of the difficulty of successive evaluation test sets are provided when a site chooses to run a common system on the data for both years. NIST is always appreciative of sites that do this.

4.1.4. Variants

Several outgrowths of the original basic one-speaker detection task have developed over time and have been evaluated.

In 2000 and 2001 the opportunity was offered to evaluate on data in a language other than English. Similar task protocols were implemented using the (non-conversational) Spanish language AHUMADA corpus [8]. NIST remains interested in obtaining access to new sources of conversational telephone speech data, particularly in languages other than English, for use in future evaluations.

As shown in figure 1, in the year 2001 we began using corpora consisting primarily of cellular data. Due to the nature of the data collection paradigm, this also meant that evaluation target trials became primarily trials involving the same telephone handset in the training and test data. This will change, however, with the new MIXER corpus first being used in the 2004 evaluation, as described in the next section.

The 2002 evaluation included a "multi-modal" (multi-channel might have been a better term) track to the one-speaker detection task using the "FBI Voice Database for Automated Speaker Recognition Systems" (described in [9]). This track involved recordings made using several different microphones and a telephone, and also introduced a no-decision option as a possible actual decision.

4.2. One-Speaker Detection with Extended Data

In 2001 some investigatory work by Doddington [10] led to the introduction of the extended data sub-task of the NIST evaluation. Increased speech content in the test segments and, especially, in the training data allowed the use of idiolectal, prosodic, linguistic, and other information sources beyond the purely acoustic information that has dominated the limited data speaker detection work. These techniques were further investigated during a workshop at Johns Hopkins University during the summer of 2002 [11].

The realization of how much system performance could be enhanced under the extended data protocols was a major milestone in the NIST Speaker Recognition evaluation series. The first implementation in 2001 was a dry run and relied on using the human generated official LDC transcripts of the Switchboard-1 data. In 2002 and 2003 regular evaluation rules applied, and all the auxiliary types of information systems were allowed to use, in particular (errorful) word-level transcripts of the conversational data, were generated by automatic systems.

4.2.1. Training Data

Instead of limiting the training data to two minutes of speech, a number of entire conversation sides were provided to build each model. There were defined training conditions

consisting of 1, 2, 4, 8 or 16 conversation sides by each target speaker.

Several types of automatically generated auxiliary information were provided to the extent they were available. For the 2003 evaluation, NIST was able to make available ASR (automatic speech recognition) system word transcripts, phone-level transcripts, handset labels, base GMM-UBM (Gaussian Mixture Model – Universal Background Model) scores, speech activity marks, pitch tracks, and language models.

NIST installed and operated an instantiation of BBN's Byblos recognizer to generate the ASR transcripts. NIST's version of Byblos was designed for real-time speed, not state-of-the-art accuracy. The transcripts were estimated to have approximately a 50% word error rate.

MIT Lincoln Labs, using a GMM-based handset classifier [12], automatically generated the handset labels. While the classifier is less than perfect, it is believed to do a very good job in distinguishing between electret and carbon-button microphones in telephone handsets. MIT Lincoln labs also provided base GMM-UBM scores and speech activity marks.

SRI provided pitch tracks for each conversation side, and the Air Force Research Laboratory provided language model probabilities [13].

4.2.2. Test Data

An entire conversation side was used as a test segment. Both the audio and the automatically created auxiliary information types, as described in section 4.2.1, were made available to each evaluation system.

4.2.3. History – State-of-the-Art

The extended data condition was offered as a dry run only in 2001. The official, human generated, transcripts were made available and developers could, if they wished, process the data, look at the results, tune their systems to them, and then reprocess the data. We therefore choose not to compare the 2001 results to those of subsequent years.

Figure 2 shows the best performance achieved in 2002 and 2003 with the eight conversation sides training condition. Thec onfidence boxes provide some indication of the significance of the best system performance gains between years, with the equal error rate decreasing from around 2% in 2002 to around 1.3% in 2003. The performance differences that these results represent over those for the limited data condition are quite remarkable.

4.3. Two Speaker Detection with Limited Data

Since conversational speech is sometimes available only in summed channel form, another variant of the basic one-speaker task that has been included in the NIST evaluations is multi-speaker detection. The challenge here is to determine whether or not a specified target speaker is present in a test segment with two (or possibly more) speakers speaking. This sub-task was implemented first in the 1999 evaluation, using summed channel data with two conversing speakers.

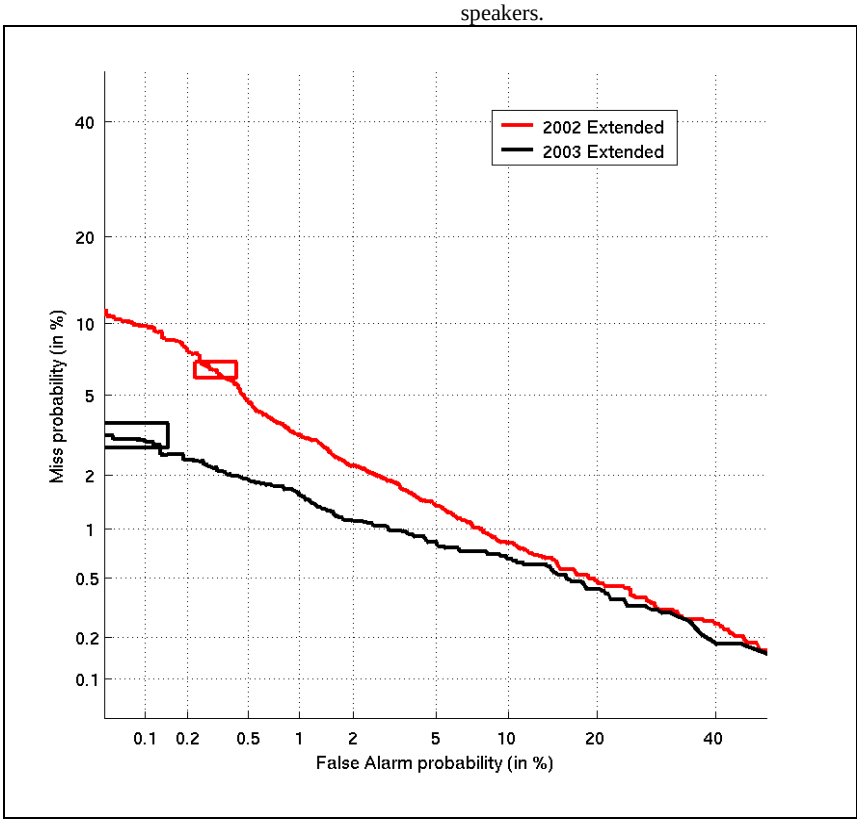

Figure 2: Best system performance history for one-speaker detection with extended data. These DET Curves are for training consisting of eight conversation sides for each target speaker.

4.3.1. *Training Data*

Originally the two-speaker detection sub-task involved two-speaker test segment data only, and used the same training data as was used for one-speaker detection. In 2003, however, two-speaker training data was also introduced. In this case, in place of the standard two minutes of speech data from one speaker, systems were to build speaker models from three whole summed channel conversations of about five minutes each. The target speaker was guaranteed to be present in each of the three conversations and the "other" conversants were guaranteed to be different in all of them.

4.3.2. *Test Data*

The test data segments for the two-speaker detection task were summed channel continuous segments with a duration of approximately one minute. Some segments involved two same sex speakers, while others had opposite sex speakers.

4.3.3. *History – State-of-the-Art*

The two-speaker detection task has been part of NIST Speaker Recognition evaluations since 1999. Figure 3 shows DET curves of the best system performance each year for test segments with two same sex speakers.

As with one speaker detection with limited data, performance was largely better with landline data (solid lines) than with cellular data (broken lines). But the curves shown suggest year to year improvements over the three years of similar test conditions using landline data and the two years, again with similar test conditions, using cellular data. Table 2 lists the equal error rates for each of these years.

Year	Equal Error Rate
1999 Landline	13.7%
2000 Landline	13.1%
2001 Landline	12.8%
2002 Cellular	16.3%
2003 Cellular	15.6%

Table 2: Equal error rates for the best system of each year's two-speaker detection task, using same sex segments.

4.3.4. *Variants*

As noted above, in 2003 summed channel training data for this condition was offered as well. Mixed mode tests, with one-speaker training and two-speaker test data, or vice versa, were additional options.

5. 2004 Evaluation

The 2004 NIST Speaker Recognition evaluation took place in the spring of 2004 utilizing data from a new LDC collection effort. This evaluation involved speaker detection as its only task, and sought to unify the previously separate efforts in recognition involving limited data or extended data and involving one-speaker or two-speaker detection.

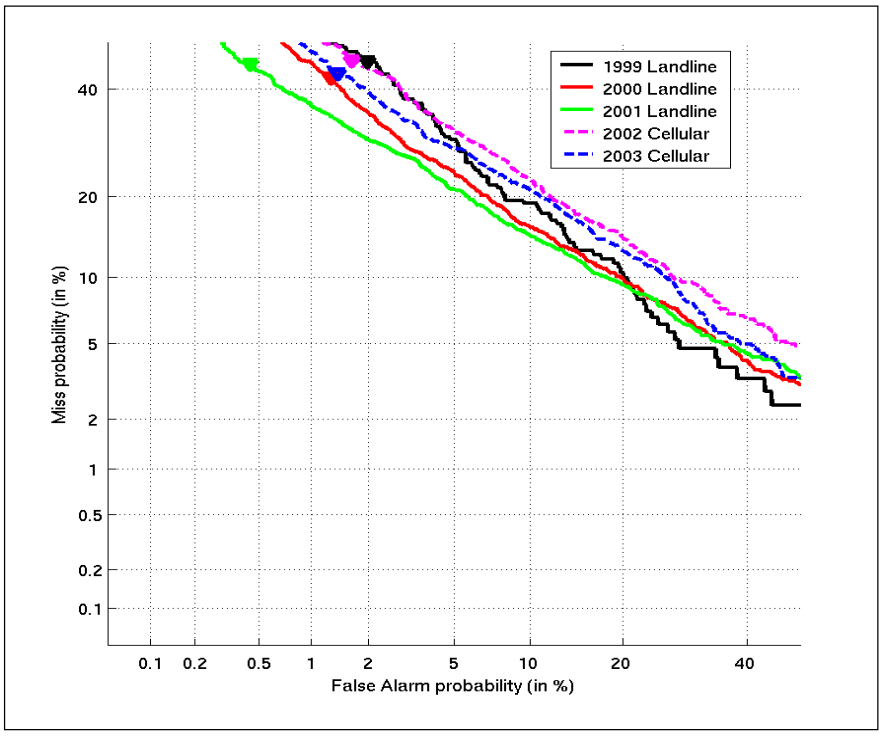

Figure 3: Best system performance history for two-speaker detection with limited data for trials involving test segments with two speakers of the same sex.

5.1. Data Collection

A new corpus collection (see [14], [15]) denoted MIXER, is being undertaken by the Linguistic Data Consortium to support this and future NIST evaluations. Multiple calls of six minutes duration on an assigned topic are being sought from each participant, with many callers encouraged to take part in as many as 25 conversations. The LDC platform calls out to participants pre-specified phone numbers on days and at times they indicated availability, but these speakers are also encouraged to initiate calls to the platform using different unique phone numbers and, presumably, unique telephone handsets. As of late March 2004 some 8500 conversations involving over 3400 participating talkers had been collected.

A special effort is being made to recruit bilingual subjects who speak Arabic, Mandarin, Russian or Spanish in addition to English. These speakers are paired with a speaker of the same language when one can be found. Thus conversations in two different languages are collected for such speakers. This will support the study of the effect of language, particularly differences between training and test language, on speaker recognition performance.

In each conversation each speaker is asked to specify the type of telephone transmission involved (cellular, cordless, or regular landline) and the type of handset used (speakerphone, headset, ear-bud, or hand-held). The callers also provide in their program registration information on their place of birth, age, and level of education. This will support the subsequent study of these factors on performance.

5.2. Test Conditions

Using the newly available MIXER data, it was decided to unify in format for this evaluation the previously separately defined one and two speaker and limited and extend data detection tasks. All training data for each target speaker was selected to come from a different phone number, and presumably a different telephone handset, from all test segment data. Seven different training conditions and four different test segment conditions were defined. The seven training conditions are summarized in table 3, while the four test segment conditions are summarized in table 4.

Tests for all 28 possible combined conditions were offered (see table 5). Participating systems could do as many or few of these as the chose, with a single required core test specified. This test uses one conversation side (of five minutes duration) as training and one such side as the test segment data. Systems undertaking multiple tests will allow study of the effects of the different training and test segment conditions on performance.

An unsupervised adaptation condition was offered for the first time in this evaluation. For each target speaker model, the trials involving it could be processed in order, and the test segments of each trial could optionally be used to modify the model as used in subsequent trials. This adaptation had to be done without knowing whether or not the test segment contained the target speaker (making the trial a target trial).

Training Condition	Description
16 sides	16 conversations sides, each consisting of a 5-minute excerpt from a full 6-minute call. When possible, they involve a single handset and language.
8 sides	8 conversation sides. When possible, they are subsets of 16-sides models.
3 sides	3 conversation sides. When possible, they are subsets of 8-sides models.
1 side	1 conversation side. Always taken from a 3-sides model.
30 seconds	A variable length segment containing about 30 seconds of speech. Each segment is taken from a corresponding 1-side model.
10 seconds	A variable length segment containing about 10 seconds of speech. Each segment is taken from a corresponding 30 seconds model.
3 conversations	3 summed-channel conversations. In general, the conversations include the sides of a 3-sides model.

Table 3: Training conditions defined for the 2004 NIST evaluation.

Test segment condition	Description
1 side	A full five minute segment from a conversation side.
30 seconds	A variable length segment containing about 30 seconds of speech. Each segment is taken from a corresponding 1-side test segment.
10 seconds	A variable length segment containing about 10 seconds of speech. Each segment is taken from a corresponding 30 second segment.
1 conversation	1 summed-channel conversation, one or both sides of which are 1 side test segments.

Table 4: Test segment conditions defined for the 2004 NIST evaluation.

Unsupervised adaptation was an available option for each of the 28 tests.. Results without adaptation were also required, permitting analysis of the performance effects of such adaptation..

		Test Segment Condition			
		10 sec	30 sec	1 side	1 conv
Training Condition	10 sec	X	X	X	X
	30 sec	X	X	X	X
	1 side	X	X	X	X
	3 sides	X	X	X	X
	8 sides	X	X	X	X
	16 sides	X	X	X	X
	3 convs	X	X	X	X

Table 5: Matrix of training and test segment conditions. The shaded entry is the required core test condition.

5.3. Evaluation Data

The test data was selected from the MIXER conversations available early in the year. In all 310 speakers were selected as target speakers. As shown in Table 6, most of these were bilingual, allowing investigation of the effect of language on performance.

Other Language	Speakers
Arabic	52
Mandarin	46
Russian	48
Spanish	79
English only	85
Total	310

Table 6: Target speakers included in the 2004 evaluation data by language spoken in addition to English.

Table 7 shows the numbers of conversation sides included in the evaluation training and test segment data, by language actually spoken. Though most of the speakers are bilingual, it may be seen that English is the predominant language of the data.

Language	Training Sides	Test Sides
English	2515	907
Arabic	300	96
Mandarin	238	64
Russian	274	61
Spanish	99	48

Table 7: Numbers of conversation sides, by language being spoken, included as training or test segment data in the 2004 evaluation.

Table 8 lists the numbers of target speakers, models (multiple per speaker in some cases), and target and non-target trials for the different training conditions in the various evaluation tests. The numbers of speakers with sufficient conversations for 16-side training was more limited than for the other training conditions.

Tables 9 and 10 show the distributions of the transmission type and handset type for the conversation sides of the core test, as reported by the talkers. The data should be sufficient to obtain meaningful results on how cellular, cordless, or plain landline transmission, and their match or mismatch between training and test affect performance. The results on the effects of different handset types may be less clear.

Model Type	Speakers	Models	Target Trials	Impostor Trials
16 sides	121	123	470	4594
8 sides	307	398	1498	15482
3 sides	310	458	1778	17703
3 convs	309	538	2068	20880
1 side	310	417	2392	23832

Table 8: For each model type, numbers of target speakers, individual models, target trials, and non-target (impostor) trials in each test. The figures for the 10 and 30 second model types are identical to those for the 1 side type. (The figures on trials given apply for the three single channel test segment types, and are slightly different for the summed channel single conversation test segments.)

Type of Transmission	Training Sides	Test Sides
Landline	257	580
Cellular	178	361
Cordless	176	219
Other/unknown	5	16

Table 9: Phone transmission types of the training and test conversation sides for the core test condition included in the NIST 2004 evaluation data.

Type of Handset	Training Sides	Test Sides
Speakerphone	37	67
Headset	107	116
Ear-bud	42	63
Regular (hand-held)	452	914
Other/unknown	5	16

Table 10: Phone handset types of the training and test conversation sides for the core test condition included in the NIST 2004 evaluation data.

The extended data tests of previous evaluations provided (errorful) word transcripts of the speech data generated by an ASR system. For the past two years this has been a real-time system that is far from state-of-the-art in its error rate. This year BBN agreed to run its relatively fast (10-20

times real-time) state-of-the-art system (similar to that described in [16]) on all the evaluation transcripts. The resulting word-level transcripts were supplied to participants along with all the training and test segments used in each of the evaluation tests.

5.4. Participants

There were twenty-five registered participating sites in the evaluation. They included research labs from companies, non-profit organizations, governments, and universities in the United States, the United Kingdom, Spain, the Netherlands, France, Switzerland, Greece, South Africa, Israel, India, China, and Australia.

5.5. Results

Full evaluation results will be presented at the NIST Evaluation Workshop in Toledo, Spain, in June 2004.

Summary results of the best performance achieved in the evaluation will be presented at the main Odyssey Speaker and Language Recognition Workshop immediately preceding the NIST workshop. This presentation will also include analysis of the effects on performance of various factors including language, telephone transmission type, and handset type.

6. Future Evaluations

The NIST evaluations are expected to continue in future years. Additional MIXER collection data, from speakers other than those selected in 2004, will be available for use.

The MIXER data collection also will include some multi-channel data that was not available in time for the 2004 evaluation. At several sites a limited number of speakers were recruited to take part in general MIXER conversations, but while in a room with a custom designed recording system that would simultaneously record their voices on eight channels. These would include two cell phone handsets, a Dictaphone, and five microphone types resembling ones in courtrooms or interview rooms. Results comparing performance for these different channels in future evaluations are expected to be of interest to agencies involved in forensic applications of speaker recognition.

It should be noted that the NIST evaluations are open to all who find the task of interest and wish to participate and report on their systems at the follow-up evaluation workshops. They are designed to be simple to implement, to be accessible to those wanting to participate, and to focus on the core issues of speaker recognition technology.

7. References

[1] Martin, A., and Przybocki, M., "The NIST Speaker Recognition Evaluation Series", *National Institute of Standards and Technology's web-site*, http://www.nist.gov/speech/tests/spk

[2] Martin, A., et al., "NIST Language Technology Evaluation Cookbook", *Proc. LREC '04*

[3] Swets, J., ed., "Signal Detection and recognition by Human Observers", John Wiley & Sons, Inc., pp. 611-648, 1964

[4] Martin, A., et al., "The DET Curve in Assessment of Detection Task Performance", *Proc. Eurospeech '97*, Vol. 4, pp. 1899-1903

[5] Martin, A., et al., "The 1996 NIST Speaker Recognition Evaluation Plan", *National Institute of Standards and Technology's*, ftp://jaguar.ncsl.nist.gov/evaluations/speaker/1996/plans/Spkr_Rec.04.v3.ps

[6] Linguistic Data Consortium, "Catalogue of Speaker Recognition Corpora", http://www.ldc.upenn.edu/Catalog/SID.html

[7] Przybocki, M., and Martin, A., "NIST's Assessment of Text Independent Speaker Recognition Performance", *COST 275 Workshop – Thee Advent of Biometrics on the Internet*, pp. 25-32

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

[9] Nakasone, H. and Beck, S., "Forensic Automatic Speaker Recognition", *Proc. 2001: A Speaker Odyssey*, The Speaker Recognition Workshop, Crete, Greece, June 18-22, 2001, pp. 139-144

[10] Doddington, G., "Speaker Recognition based on Idiolectal Differences between Speakers, *Proc. Eurospeech '01*, Vol. 4, pp. 2521-2524

[11] "SuperSID: Exploiting High-Level Information for High-Performance Speaker Recognition", The Center for Language and Speech processing, *2002 Summer Workshop*, http://www.clsp.jhu.edu/ws2002/groups/supersid/

[12] Quatieri, T., et al., "Magnitude-Only Estimation of Handset Nonlinearity with Application to Speaker Recognition", *Proc. ICASSP '98*, pp. 745-748

[13] Clarkson, P., and Rosenfeld, R., "Statistical Language Modeling using the CMU-Cambridge Toolkit", *Proc. Eurospeech '97*, Vol. 5 pp. 2707-2711

[14] Campbell, J., et al., "The MMSR Bilingual and Crosschannel Corpora for Speaker Recognition Research and Evaluation", *Proc. Odyssey '04*

[15] Martin, A., et al., "Conversational Telephone Speech Corpus Collection for the NIST Speaker Recognition Evaluation 2004". *Proc LREC 2004*

[16] Schwartz, R., et al., "Speech Recognition in Multiple Languages and Domains: The 2003 BBN/LIMSI EARS System", *Proc. ICASSP 2004*