NIST Speaker Recognition Evaluations Utilizing the Mixer Corpora—2004, 2005, 2006

Mark A. Przybocki, Alvin F. Martin, and Audrey N. Le

***Abstract*—NIST has coordinated annual evaluations of text-independent speaker recognition from 1996 to 2006. This paper discusses the last three of these, which utilized conversational speech data from the Mixer Corpora recently collected by the Linguistic Data Consortium. We review the evaluation procedures, the matrix of test conditions included, and the performance trends observed. While most of the data is collected over telephone channels, one multichannel test condition utilizes a subset of Mixer conversations recorded simultaneously over multiple microphone channels and a telephone line. The corpus also includes some non-English conversations involving bilingual speakers, allowing an examination of the effect of language on performance results. On the various test conditions involving English language conversational telephone data, considerable performance gains are observed over the past three years.**

***Index Terms*—Cross-channel evaluation, decision error tradeoff (DET) curves, Mixer Corpora, NIST evaluations, speaker recognition evaluation.**

I. 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]–[5]. During the eleven years of NIST Speaker Recognition evaluations, the basic task of speaker detection, determining whether or not a specified target speaker is speaking in a given test speech segment, has been the primary evaluation focus. This task has been posed primarily utilizing various telephone speech corpora as the source of evaluation data.

By providing explicit evaluation plans, common test sets, standard measurements of error, and a forum for participants to openly discuss algorithm successes and failures (see [6]), the NIST series of Speaker Recognition Evaluations (SREs) [7] has provided a means for chronicling progress in text-independent speaker recognition technologies.

As noted above, we have previously discussed the earlier history of the NIST SREs (for example, in [1]). Here, we concentrate on the evaluations of the past three years (2004–2006). These recent evaluations have been distinguished notably by the use of the Mixer Corpora of conversational telephone speech as the primary data source, and by offering a wide range of (mostly optional) test conditions for the durations of the training and test data used for each trial. One test condition in the past two years has involved the use of the Mixer data simultaneously recorded over several microphone channels and a telephone line. In addition, the corpus and the recent evaluations have included some conversations involving bilingual speakers speaking a language other than English. The use of Mixer data has made the recent evaluations larger and richer in the range of performance factors available for study. We discuss a few of these here.

Manuscript received February 20, 2007; revised June 5, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Joseph Campbell.

The authors are with the Speech Group, Information Access Division, Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899-1070 USA (e-mail: mark.przybocki@nist.gov; alvin.martin@nist.gov; audrey.le@nist.gov).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2007.902489

II. Evaluation Measures

An evaluation test consists of a series of *trials*, in each of which the system must determine whether a given speaker, whose *model* is defined by specified training speech data, is speaking in a given test segment. Test trials can be categorized as either *target trials*, meaning the target speaker is speaking in the test segment (correct answer is true), or *impostor* (nontarget) *trials*, meaning the target speaker is not speaking in the test segment (correct answer is false). Each trial requires two outputs from the system under test, namely an *actual decision*, which declares whether or not the test segment contains the specified speaker, and a numeric *likelihood score*, which quantifies the system's degree of belief that the target is speaking. (Larger scores imply greater likelihood of this.) There are two types of actual decision errors, *missed detections* (target trials) and *false alarms* (impostor trials). The *miss rate* ($P_{\mathrm{Miss|Target}}$) is the percentage of target trials decided incorrectly (as false). The *false alarm rate* ($P_{\mathrm{FA|Impostor}}$) is the percentage of impostor trials decided incorrectly (as true).

A. 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. The weights depend on the assumed costs of a missed detection and of a false alarm, and on the assumed *a priori* probability of a target trial. We then define equation (1), shown at the bottom of the next page. The parameters here are inherently application specific. 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 probability need not, and does not, correspond to the actual target richness of the evaluation data trials but rather reflects application scenarios of possible interest, as do the cost parameters specified.

The cost function is made more intuitive by normalizing it so that a system with no discriminative capability is assigned a cost of 1.0. Since (1) implies that deciding "false" for every trial results in a numerator of 0.1, while deciding "true" for every trial

results in a numerator of 0.99, we set NormFact to the minimum of these two values, namely 0.1.

B. Decision Error Tradeoff (DET) Curves

In addition to determining single number measures such as C_{DET} cost, more performance information can be shown in a graph plotting all the possible operating points of a system based on the likelihood scores. By sweeping over all possible likelihood values as thresholds for separating decisions of true and false, all possible system operating points are generated.

NIST has used a variant of the popular receiver operating characteristic (ROC) curve, which has long been utilized to represent decision task performance by Swets [8] and others, where the two error rates are plotted on the x and y axes on a normal deviate scale. NIST introduced the use of DET curves [9] in the 1996 evaluation [10], and DET curves have since been widely used for the representation of detection task performance.

The actual decision point on the DET curve, corresponding to C_{DET} value, can be marked with a special symbol for easy identification. A confidence box may be drawn around this point corresponding to 95% confidence limits for the miss and false alarm rates. (This assumes trial independence, which is not fully valid.) The point on the curve correspond to the minimum possible C_{DET} value can also be marked. The distance between the minimum and actual C_{DET} points indicates how well the actual decision threshold is calibrated. (Another popular measure is the equal error rate, defined as either coordinate of the intersection of the DET curve with the line $x = y$.)

Fig. 1 shows the DET curves for 36 primary systems for the "core" test (to be discussed in Section IV) in the 2006 evaluation. The best performing systems are highlighted. Note that most of these curves are close to linear, as would be implied by normality in the underlying distributions of target and nontarget trial scores.

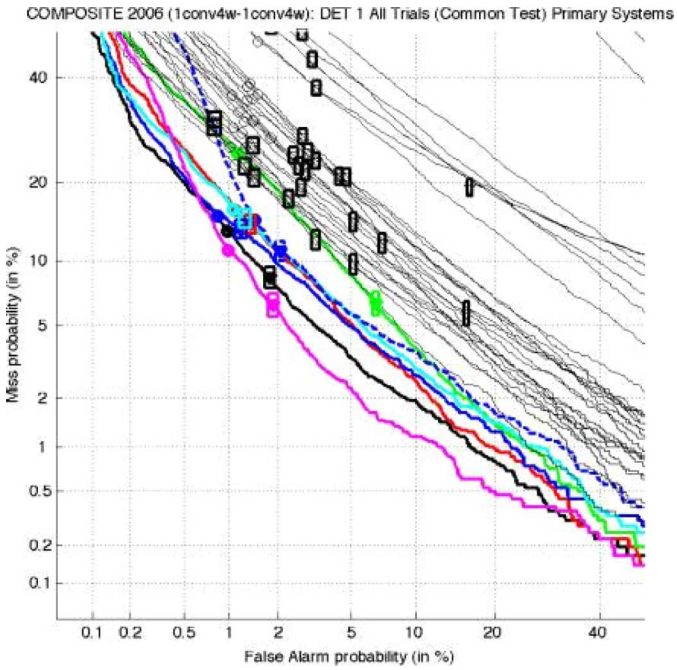

Fig. 1. DET curves for the "core" test condition for 36 primary systems in the 2006 evaluation.

III. Mixer Corpora

Appropriate data is essential for research in speaker recognition, and large quantifies of appropriate data are needed for significant evaluation. NIST has benefited from the ongoing collections of conversational telephone speech by the Linguistic Data Consortium [11]. Several collections of Switchboard style corpora [12], each of which included hundreds of speakers and thousands of conversations, were used extensively in the detection tasks of the NIST Speaker Recognition Evaluations from 1996 to 2003.

The 2004, 2005, and 2006 evaluations all used conversational speech data of the recently collected Mixer Corpora of the LDC. These corpora are based on a platform utilizing an automaton that can initiate contacts via phone to find pairings of registered participants to engage in recorded conversations on assigned topics. As with the previously used Switchboard platform, the participants can also initiate calls, and have the platform find them a conversational partner. The objective is to secure from a large number of target speakers a significant number (eight or more for the recent evaluations) of conversation sides from a single handset (telephone number)[1] that may be used for training, and some number of conversations from other handsets, which may be used for test segment data. The compensation paid to the registered participants includes incentives to accomplish this. See [13]–[15].

IV. Evaluation Conditions

The Mixer collection was initiated following earlier NIST evaluations and other research [16]–[18] suggesting that considerable performance benefits could be achieved with longer durations of training and, to a lesser extent, test segment data. It was desired, therefore, to have a range of training and test duration conditions, with up to eight conversation sides (averaging about 2.5 min of speech each) for training and a single full conversation side for test. On the other hand, there was considerable interest among some participants in having very short duration training and test conditions, namely 10 s each, reflecting considerable demand for such in commercial applications. Thus a matrix of possible combinations for training and test was suggested.

[1]The imperfect assumption is made that different handsets correspond to different phone numbers.

$$C_{\mathrm{DET}} = \frac{((C_{\mathrm{Miss}} * \mathrm{P}_{\mathrm{Miss|Target}} * \mathrm{P}_{\mathrm{Target}}) + (C_{\mathrm{FA}} * \mathrm{P}_{\mathrm{FA|Impostor}} * (1 - \mathrm{P}_{\mathrm{Target}})))}{\mathrm{NormFact}} \quad (1)$$

TABLE I
TRAINING CONDITIONS IN 2005 AND 2006

Training Condition	Description
10 seconds (of one side)	A variable length segment containing about 10 seconds of speech. Each is a sub-segment of the data for a 1 side model.
1 side (average 2.5 min of speech)	1 conversation side.
3 sides (average 7.5 min of speech)	3 conversation sides, generally a subset of the sides of an 8 sides model.
8 sides (average 20 min of speech)	8 conversation sides.
3 conversations	3 summed-channel conversations. In general, the conversations include the sides of a 3 sides model.

The advantage of longer duration segments lay in the possibility of using various types of higher level information about the speech content of the data. One such type of information is the actual words and word combinations used, i.e., the statistics on unigrams, bigrams, etc. Since many of the participating sites did not have their own systems for automatic speech recognition (ASR), the recent evaluations have included ASR transcripts, with a word error rate on the order of 20%, produced by an up-to-date, relatively fast, system.[2, 3]

A. Training Conditions

Table I lists the five training condition in the 2005 and 2006 evaluations. (In 2004, there were also a 30-s training condition, and a 16-sides condition; unfortunately, data for the latter was limited.) The first four reflect the desired range of speech durations provided for the target speaker models. The final condition reflects the desire to include summed channel (sometimes called 2-wire) data in the evaluation. Three such conversations are provided, with the target speaker guaranteed to be one of the two speakers in each, while the three other speakers are all distinct. Systems must automatically segment the training speech data to find that corresponding to the target.

B. Test Segment Conditions

Table II lists four the test segment conditions for the 2005 and 2006 evaluations. (The 2004 evaluation also had a 30-s condition, but not a microphone condition.) The first two represent the two extremes for single conversation side telephone data, namely 10 s of speech or the whole conversation side. The third is the summed channel conversation side (2-wire) condition, where the system must determine if either speaker in the conversation is the target. The final condition is the microphone data for cross-channel speaker detection. A subset of the test conversations were simultaneously recorded over eight microphone channels, and each such recording is included as a test segment for this condition. The corresponding telephone versions are included in the first condition.

[2]NIST thanks BBN Technologies for producing these transcripts and making them available to evaluation participants.

[3]Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

TABLE II
TEST SEGMENT CONDITIONS IN 2005 AND 2006

Test Segment Condition	Description
10 seconds (of one side)	A variable length segment containing about 10 seconds of speech. Each is a sub-segment of a 1 side segment.
1 side (average 2.5 min of speech)	A full five minute segment from a conversation side.
1 conversation	1 summed-channel conversation, one or both sides of which are 1 side test segments.
1 microphone side	A 1 side conversation included above as recorded on one of eight auxiliary microphone channels

TABLE III
MATRIX OF TRAINING AND TEST SEGMENT CONDITIONS INCLUDED AND REQUIRED IN 2006

		Test Segment Condition			
		10 sec	**1 side**	**1 conv**	**1 mic conv**
Training Condition	**10 seconds**	optional			
	1 side	optional	required	optional	optional
	3 sides	optional	optional	optional	optional
	8 sides	optional	optional	optional	optional
	3 conversations		optional	optional	

C. Matrix of Tests

The combination of training and test segment conditions leads to a matrix of possible tests to include in the evaluation. In 2004 and 2005, the full matrix of tests was used; for 2006, it was decided to pare this matrix to those in which there was greater interest among participants. There was, for example, limited interested in short duration training with longer test segments.

Table III shows the matrix of 15 offered training and test segment conditions as specified for the 2006 evaluation. Note that the condition of training and testing on single telephone conversation sides was the "core" condition; this test was required of all participating sites. Beyond this, they could do as many, or as few, of the other tests as they preferred. However, for each test chosen, systems were required to do all trials included in each test. In particular, this meant doing all the trials that involved speech in languages other than English.

Sections VI–IX will present performance results from recent evaluations for various test conditions.

2006 Evaluation Participation

- 36 submitting sites

Australia	Canada	China (6)
Czech Republic	Denmark	Finland
France (8)	Germany (2)	Israel
Italy	Lebanon	Singapore (2)
South Africa	Spain (2)	Switzerland
United Kingdom	United States (6)	

- 96 systems
- 283 test condition/system combinations

Fig. 2. Participation in the 2006 evaluation.

V. Participants

Participation in the NIST SREs has been growing over the years. There were 24 participating sites in both 2004 and 2005, more than in any previous evaluations. Then, in 2006, the number of participants grew by 50%, with 36 different sites (or teams of sites) submitting one or more evaluation systems. The participating sites included research labs from companies, nonprofit organizations, governments, and universities in North America, Europe, the Middle East, Africa, Asia, and Australia, reflecting worldwide interest in this technology.

Fig. 2 summarizes the participation seen in 2006. The 36 sites together produced 96 different systems which collectively gave 283 sets of evaluation results across the 15 test conditions offered in 2006 as discussed in Section IV-C. Note that only six of the 36 participating sites were in the U.S. The majority were from Europe, with France as the biggest player, while East Asian participation saw large increases in 2006 and other recent years.

It should be noted that the accepted community practice, at least until now, has been not to publicize evaluation winners and losers as such by identifying participating sites with their performance results in open meetings and publications. This has been 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 recent NIST Speaker Recognition evaluations, each site agreed 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. There has been some debate about whether a more publicly open policy would be better for the larger community, and a different policy may be adopted for the next evaluation.

Following current policy, the DET plots presented in this paper show curves for all participating sites or for the best performing systems for various evaluation conditions without providing the corresponding site names.

VI. Core Test and Common Condition

The core test, as described above, consisted of all trials with a single telephone conversation side for both training and test. It is considered the central test of these evaluations, and has been required of all participants.

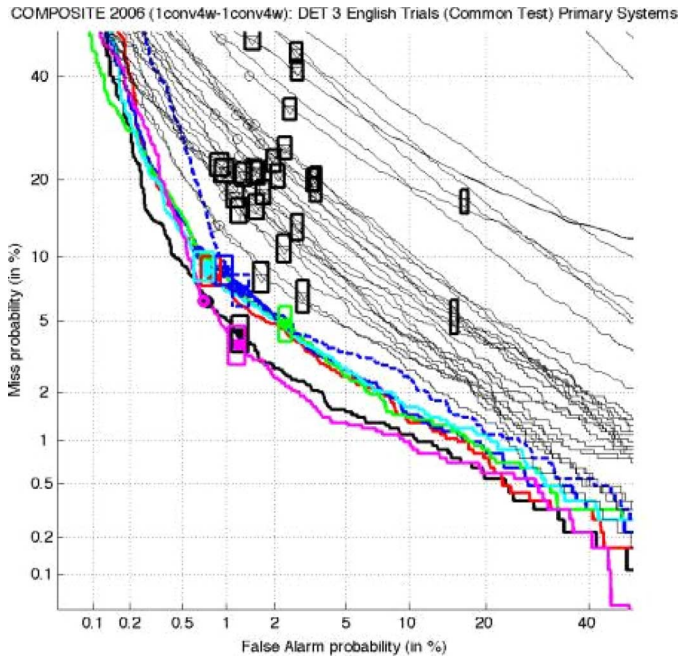

Fig. 3. Common condition results for 36 primary systems in the 2006 evaluation.

Fig. 1 (presented in Section II-B) shows the core test DET curves for the primary systems of the 36 participants in the 2006 evaluation. The several top performing curves are highlighted.

NIST has also defined a "common condition" in each evaluation, a subset of the core test consisting of the trials of greatest interest. In 2006 (and with slight differences, in the two preceding years), the common condition consisted of all core test trials in which all of the speech data, training and test, was in English. Thus, this condition filters out the effects of having multiple languages (discussed in Section X below) and is more comparable to previous results.

Fig. 3 shows the common condition DET curves of the 36 primary systems. Comparing Figs. 2 and 3, it can be noted that best systems performed better on the common condition than on the core test, particularly in the vicinity of the curves near the actual and minimum C_{DET} points (the triangles and circles). Since the evaluation emphasized this score on the common condition, this received the greatest effort by system designers, with some systems de-emphasizing the non-English trials. (Note that the ASR transcripts that were provided were English only.)

Fig. 4 shows DET curves for the common condition of the best systems in 2004, 2005, and 2006. It may be observed that a large improvement was obtained in 2005 compared to 2004. The situation in 2006 compared to 2005, however, was more equivocal, with the curves intersecting and improvement limited to lower miss rates. NIST hopes to investigate further the reasons for this.

Do the performance improvements seen in Fig. 4 (and in plots to be presented below) reflect real algorithm improvements, or could they result simply from changes in the evaluation data? One advantage of using a consistent corpus type such as Mixer over a number of evaluations is that it should minimize data differences between evaluations, particularly when restricted to a common test condition, including only English conversations.

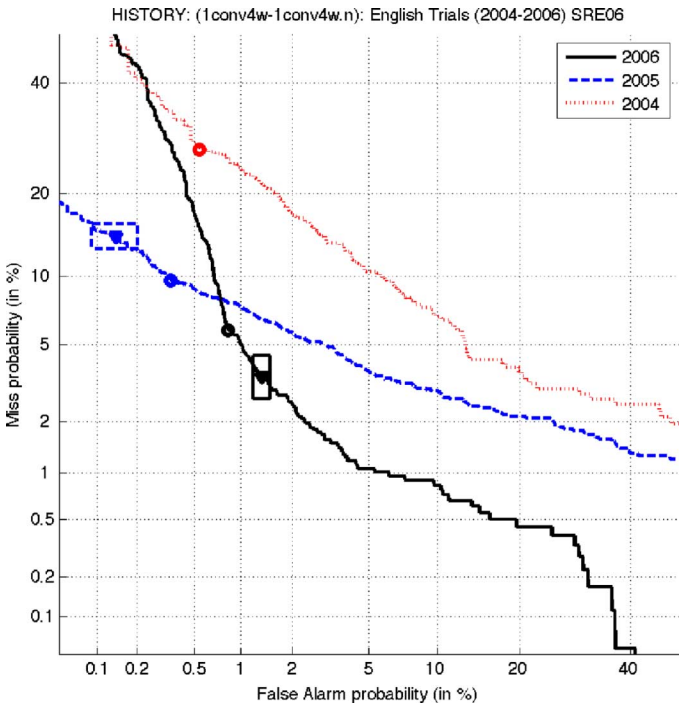

Fig. 4. Common Condition best performing systems results 2004–2006.

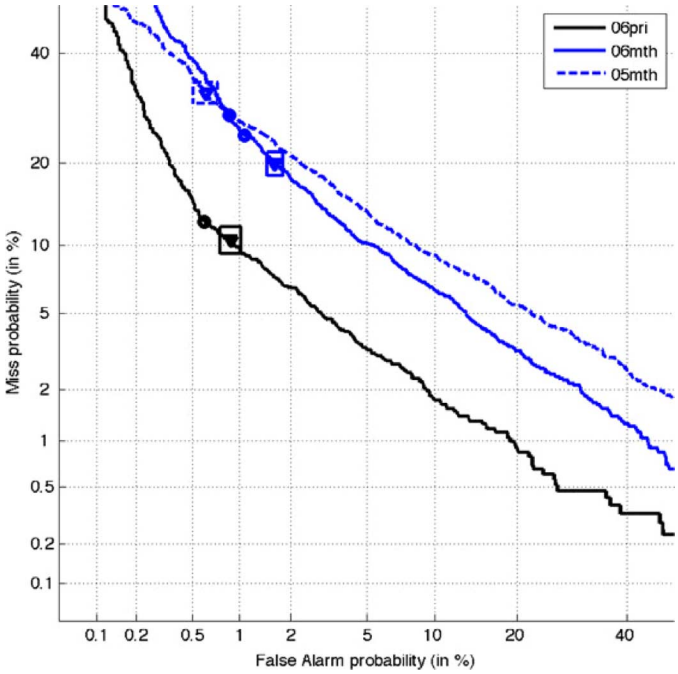

Fig. 5. Mothballed system performance for the common condition for one site on the 2005 and 2006 evaluation test sets, along with the site's primary system performance on the 2006 data.

However, assuring a constant level of difficulty for evaluation data can be a problem even when consistent data collection protocols are employed. This problem has been observed in several of the other NIST speech processing evaluations.

In 2005 and 2006, NIST invited sites to save "mothballed" systems from the previous evaluation, and to run them on the current year's data. Several sites responded to this request, and it was possible to compare results on the data of successive years run on common systems. Fig. 5 shows results for one such site. It can be seen that the mothballed system's results on the 2005 and 2006 test sets were not very different (the DET curves intersect), while this site's primary system showed considerably improved performance on the 2006 data. Thus, we believe that there was little difference in test set difficulty for the common condition for these two years, and that the observed performance improvements in 2006 were real.

There was one notable evaluation protocol change adopted in 2005, however. Previously, the data supplied with each trial for the common condition (and other single-channel test conditions) consisted of just the single conversation side, in both training and test. Starting in 2005, both sides of each conversation were made available, separately, with the side of interest designated. This additional data could assist in modeling the nature of the conversation taking place. Only a few participating sites have attempted to make use of this additional information, however. It may have contributed to the improved best system performance of 2005 compared with 2004 but was probably not the major reason for the improvement.

VII. Extended Training

As noted previously, for the past several years the NIST evaluation has included an extended training data component in its evaluation because of the enhanced performance results this has

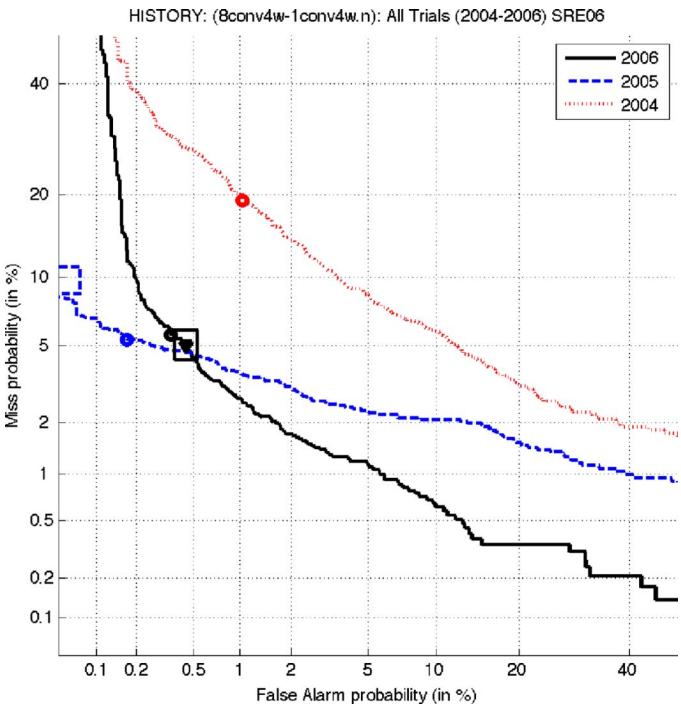

Fig. 6. Extended training condition best performing system results 2004–2006.

offered. This has concentrated on the test condition where eight training conversation sides are provided for each target speaker (generally using a common handset), along with a single conversation side for each test segment. Fig. 6 shows the best system performance for this condition over the past three evaluations.

As with the common condition, there was large performance improvement in 2005 over 2004, particularly in the low false alarm region of the DET plots, the region of greatest interest for the C_{DET} cost function. However, there likewise was a more

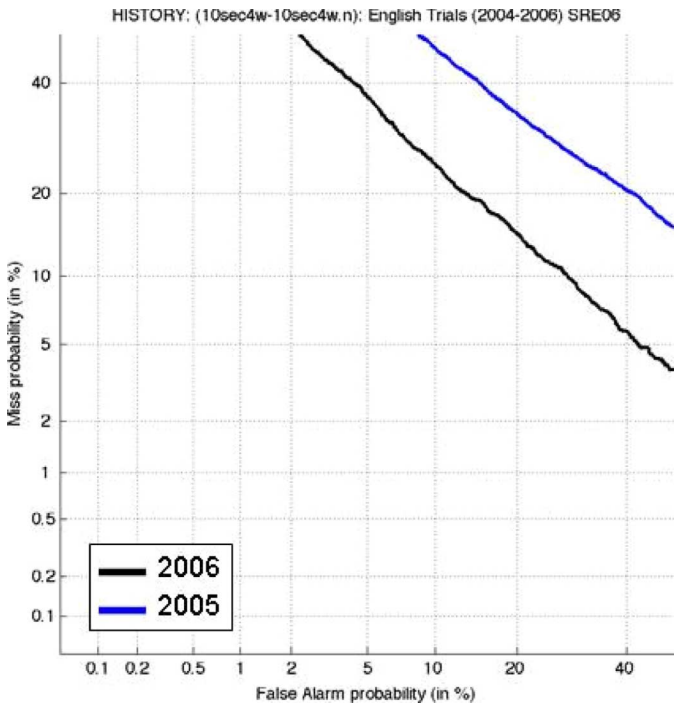

Fig. 7. Best system performance on 10-s training and test trials in the 2005 and 2006 NIST evaluations.

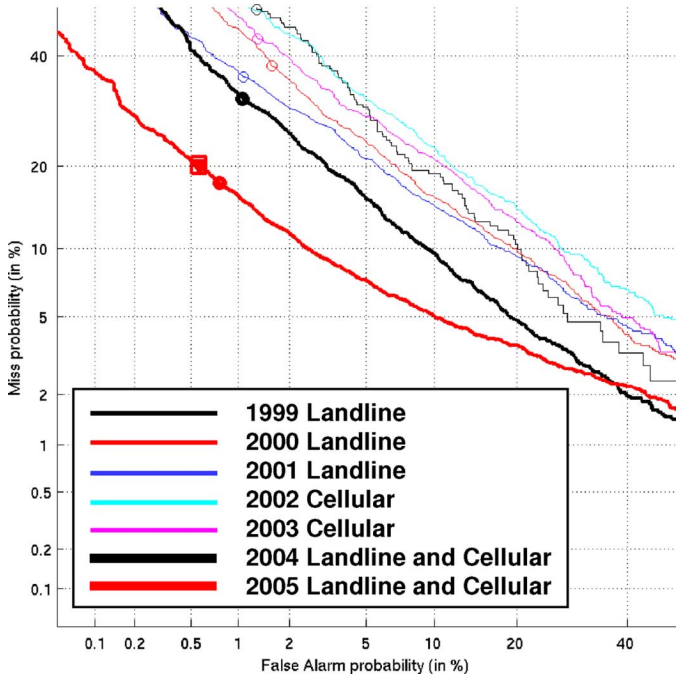

Fig. 8. Best system performance for two speaker detection 1999–2005 evaluation.

equivocal situation in 2006 compared with 2005, with the best system curves intersecting. It is also instructive to compare the curves of Fig. 6 with those of Fig. 4 to observe the improved performance afforded by the additional training data for each target speaker.

VIII. 10-s Durations

The extremely short duration (10 s for both training and test segment) test condition has, as noted previously, been maintained as part of the NIST evaluations because of participant interest and its relevance to commercial applications. Fig. 7 shows the DET curves for the best performing systems for the past two evaluations. The considerable performance improvement seen in 2006 compared with 2005 is encouraging, but most notable perhaps is the remaining large performance gap compared with other longer duration test conditions.

IX. Two-Speaker Detection

For some years, the NIST evaluations have included a two-speaker detection condition. Here, the data, training, and test is the result of summing the two channels of the phone conversations, where the target speaker participates in all training conversations, and the task is to determine whether either of the two test segment speakers is this target.

The two-speaker training has consisted of three conversations, each with the target of interest as one participant, and with three different speakers as the other participant. It is part of the task to track the speech of the single target of interest in the three training conversations, and then to find any speech of this target in the test segment, consisting of a single summed channel conversation.

Fig. 8 shows the history of best system performance results on this test for each year from 1999 through 2005. (2006 has not yet been added to the chart.) For earlier years, the data was either all landline or all cellular, while since 2004 the Mixer Corpora have provided as mix of the two. General progress is apparent, with a setback in performance seen in 2002 and 2003, when the data switched from landline to cellular. Also, in the three earlier years, only the test segments consisted of summed channel data; the training was single channel. Sizable performance improvements are seen for the two most recent years shown using Mixer data. Fig. 8 may be compared with Fig. 4 to get a sense of the gap between one-speaker and two-speaker performance. This gap has lessened in recent years.

X. Language Effects

The Mixer Corpora have included hundreds of bilingual speakers, people who can converse fluently on assigned topics in either English or another language. Large numbers of speakers of Arabic, Mandarin, Russian, and Spanish have been recruited, as well as smaller numbers of other language speakers. The collection protocol has been designed to pair common speakers of a language other than English. This has included calling all available speakers of each non-English language at roughly the same time, setting aside special collection days for specific languages, and offering bonuses for completing a set number of non-English calls.

There has often been speculation about the effect of language on speaker recognition performance, and particularly of the effect of speakers switching language between training and test. The Mixer Corpora have allowed this to be investigated. Fig. 9 shows DET curves by language for one system for the core condition (one conversation training and test) of the 2006 evaluation. The four curves cover the four different combinations of

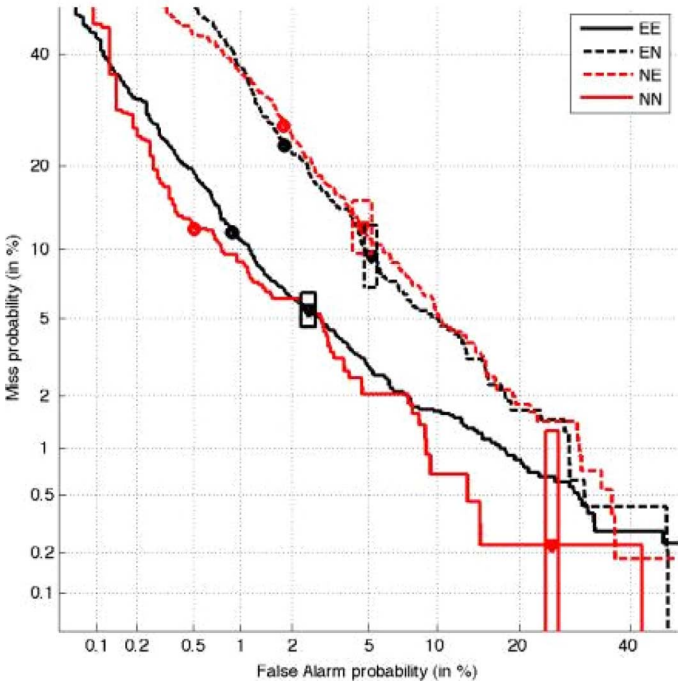

Fig. 9. DET curves for one system in the 2006 evaluation based on the language (E for English, N for non-English) in the model (training) and test segment data for all common condition trials. For example, "EN" designates English training and non-English test segments.

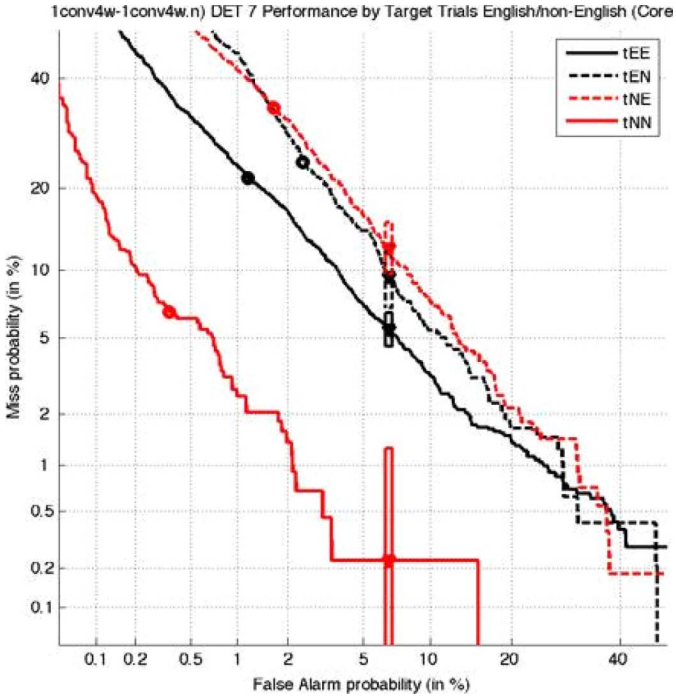

Fig. 10. DET curves for the same system as in Fig. 9 but with only the target trials limited and all nontarget trials used for each curve.

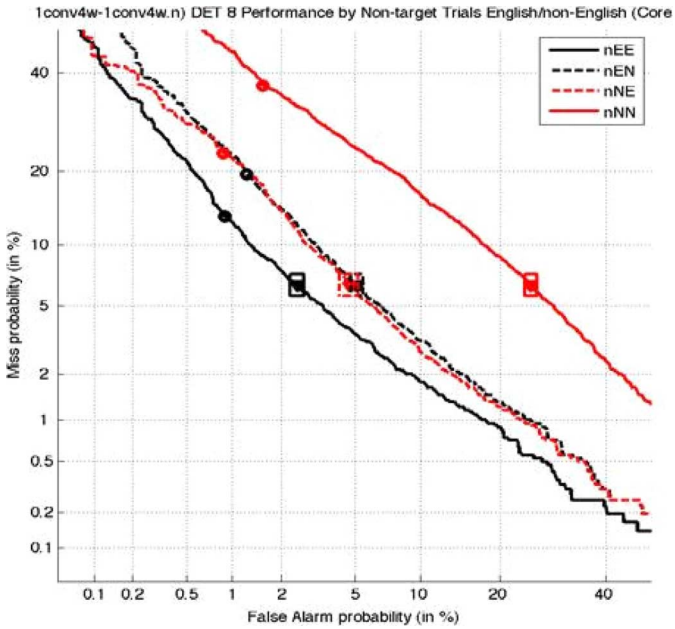

Fig. 11. DET curves for the same system as in Fig. 9 with only the nontarget trials limited and all target trials used for each curve.

English or non-English in the model (training) and test segment data for all trials (both target and nontarget trials). For example, "NE" refers to trials where the training is in a non-English language, and the test segment is in English.

It may be observed that performance is clearly superior for the matched (same language) trials than for the unmatched. Thus, language consistency matters for successful recognition, but this figure shows no significant differences depending on the identity (English or non-English) of the language(s) involved. It may be noted that the chosen (rather well performing) system is fairly typical in this regard of the many evaluation systems.

Some further insight on what is happening may be seen by looking separately at the language effects on target and nontarget trials. In Fig. 10, each DET curves includes all impostor trials, with only the target trials divided by language condition as in Fig. 9. Fig. 11 similarly divides the impostor trials, while including all target trials in each curve.

In Fig. 10, it is seen that matched target trials produce better performance, as might be expected, but the effect is far greater for the non-English target trials. Further, in Fig. 11 it is seen that matched English nontarget trials produce better performance than unmatched trials, but matched non-English trials produce far worse results. This suggest that for non-English data (conversation language information was made available) the system may be doing language recognition as much as it is doing speaker recognition. This may be a result of the non-English conversations receiving less evaluation emphasis (not being included in the common evaluation condition) and not having in-language ASR results provided for them. Interestingly, the calibration for the non-English training and non-English test conditions, indicated by the distance between the actual decision and minimal C_{DET} points, is poor in all three figures. This is a matter that should be studied further and, hopefully, future evaluation results will show some lessening of the effects observed here.

XI. Cross-Channel Performance

A key feature of the Mixer collections for the recent NIST evaluations has been the inclusion of hundreds of speakers who for several of their calls visit one of three special collection sites

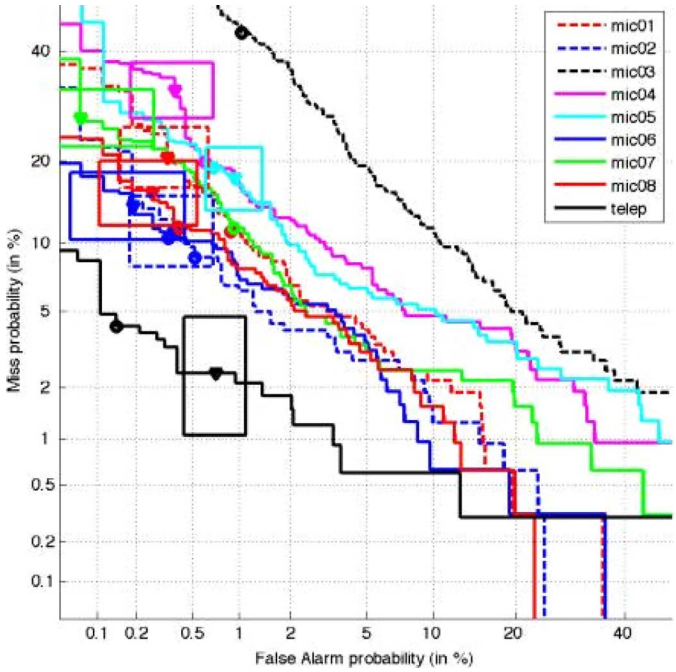

Fig. 12. DET curves one 2006 evaluation system for test segments recorded simultaneously over each of eight microphone channels and over the telephone.

where they talk using a phone but also have their side of each conversation simultaneously recorded over a collection of eight different microphones of varying types placed on or near them. One evaluation test condition during the past two years has consisted of trials where the training is regular telephone recorded data but the test segments are recorded over these different microphone channels (and also over the telephone). Fig. 12 shows DET curves for one system in the 2006 evaluation that participated in this test condition.

The relatively unsmooth curves and large confidence boxes of the figure reflect the relatively small numbers of trials in the 2006 evaluation for the cross-channel condition. The key point to note is the far better performance of the telephone test segment data (solid black) than that of all the cross-channel (all training was over the telephone) microphone curves. (The microphone used in the broken black curve was apparently defective.)

Future NIST evaluations will emphasize the cross-channel condition and will have greater amounts of data. The performance effects of the several different types of microphones will be examined in more detail, and participants interested in this problem will undoubtedly find successful approaches to reducing the performance effect of different recording conditions for training and test.

XII. Future Evaluations

After 11 years on annual speaker recognition evaluations, the NIST SRE series will go on hiatus for a year in 2007, but should resume in 2008. As noted, the size and complexity of the evaluation has grown over the years, as has the number of participating sites. The extra time to prepare for the next evaluation will allow for an enhanced collection of cross-channel data, which will be a major focus of future evaluations, and to assess the other resources needed to support continuing and expanded evaluations.

As suggested in Sections VI–IX, considerable progress has been seen in recent years in the major test conditions involving conversational telephone speech in English, which have long been the focus of the NIST evaluations. Approaches involving the merging of different types and levels of information, and various types of normalization to different telephone channel conditions have led to major improvements in the field. These are discussed in other papers in this issue. Further such progress may be expected in future evaluations, but perhaps what is most desired is improved handling of channel variability beyond the telephone domain. The handling of multiple and cross language conditions will also be of continued interest, and there remains considerable room for performance improvement when the training and test speech duration are very short.

There appears to be growing interest in the speaker recognition field and in the NIST type of evaluation among both technology developers and potential users of this technology. The Mixer collection paradigm appears to be well adapted for future efforts, but the need for ever more such data and the cost of collection remains an issue of concern.

It should be noted that the NIST evaluations remain open to all who find the task of interest and wish to participate and report on their systems at the follow-up evaluation workshops.

References

[1] M. A. Przybocki and A. F. Martin, "NIST speaker recognition evaluation chronicles," in *Proc. Odyssey 2004: Speaker Lang. Recognition Workshop*, Toledo, Spain, Jun. 2004, pp. 15–22.

[2] M. A. Przybocki, A. F. Martin, and A. N. Le, "NIST speaker recognition evaluation chronicles – Part 2," in *Proc. Odyssey 2006: Speaker Lang. Recognition Workshop*, San Juan, PR, Jun. 2006, pp. 1–6.

[3] M. A. Przybocki and A. F. Martin, "NIST's assessment of text independent speaker recognition performance," in *Proc. COST 275 Workshop—Advent Biometrics Internet*, Rome, Italy, Nov. 2002, pp. 25–32.

[4] A. F. Martin and M. A. Przybocki, "The NIST speaker recognition evaluations: 1996–2001," in *Proc 2001: A Speaker Odyssey*, Chainia, Crete, Greece, Jun. 2001, pp. 39–43.

[5] A. F. Martin, M. A. Przybocki, and J. P. Campbell, "The NIST speaker recognition evaluation program," in *Biometric Systems: Technology, Design and Performance Evaluation*, J. Wayman, Ed. *et al.* New York: Springer, 2005, ch. 8, pp. 241–262.

[6] A. F. Martin *et al.*, "NIST language technology evaluation cookbook," in *Proc. LREC '04*, Lisbon, Portugal, May/Jun. 2004, pp. 2011–2014.

[7] A. F. Martin and M. A. Przybocki, "The NIST Speaker Recognition Evaluation Series," NIST, Gaithersburg, MD [Online]. Available: http://www.nist.gov/speech/tests/spk/

[8] J. Swets, Ed., *Signal Detection and Recognition by Human Observers*. New York: Wiley, 1964, pp. 611–648.

[9] A. F. Martin *et al.*, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech'97*, Rhodes, Greece, Sep. 1997, vol. 4, pp. 1899–1903.

[10] A. F. Martin *et al.*, "The 1996 NIST Speaker Recognition Evaluation Plan," NIST, Gaithersburg, MD [Online]. Available: ftp://jaguar.ncsl.nist.gov/ evaluations/ speaker/1996/plans/Spkr_Rec.04.v3.ps

[11] "Catalogue of Speaker Recognition Corpora," Linguistic Data Consortium, Philadelphia, PA [Online]. Available: http://www.ldc.upenn.edu/Catalog/SID.html

[12] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. ICASSP'92*, San Francisco, CA, pp. 517–520.

[13] J. Campbell *et al.*, "The MMSR bilingual and crosschannel corpora for speaker recognition research and evaluation," in *Proc. Odyssey '04*, Toledo, Spain, Jun. 2004, pp. 29–32.

[14] A. F. Martin *et al.*, "Conversational telephone speech corpus collection for the NIST speaker recognition evaluation 2004," in *Proc LREC 2004*, Lisbon, Portugal, May/Jun. 2004, pp. 587–590.

[15] C. Cieri *et al.*, "The mixer and transcript reading corpora: Resources for multilingual crosschannel speaker recognition research," in *Language Resources and Evaluation Conf. (LREC)*, Genoa, Italy, May 2006, pp. 117–120.
[16] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," in *Proc. Eurospeech'01*, Aalborg, Denmark, Sep. 2001, vol. 4, pp. 2521–2524.
[17] D. Reynolds *et al.*, "The SuperSID Project: Exploiting high-level information for high-accuracy speaker recognition," in *Proc. ICASSP'03*, Hong Kong, China, Apr. 2003, pp. 784–787.
[18] "SuperSID: Exploiting high-level information for high-performance speaker recognition," Center Lang. Speech Process., Baltimore, MD [Online]. Available: http://www.clsp.jhu.edu/ws2002/groups/supersid/

Mark A. Przybocki received the M.S. degree in computer science from Hood College, Frederick, MD.

He joined the Speech Group, National Institute of Standards and Technology (NIST), Gaithersburg, MD, in 1992 and has been involved with various human language technology evaluation projects including automatic speech recognition, speaker recognition, and language recognition. Currently, he is the Project Leader for both the NIST open machine translation evaluation and the automatic content extraction evaluation while he maintains a strong interest in the area of speaker recognition as it applies to biometrics.

Alvin F. Martin received the Ph.D. degree in mathematics from Yale University, New Haven, CT, in 1977.

He has worked in the Speech Group at the National Institute of Standards and Technology (NIST), Gaithersburg, MD, since 1991. He has coordinated NIST's speaker recognition evaluations since 1996 and has been involved in various other NIST speech processing evaluations including those in language recognition and large vocabulary continuous speech recognition. He has taught mathematics and computer science at the college level and worked on the development of automatic speech recognition and speech processing systems before coming to NIST.

Audrey N. Le is a graduate of Mississippi State University.

She works in the Speech Group at the National Institute of Standards and Technology (NIST), Gaithersburg, MD. She has been involved in various HTL-related evaluations conducted by the Speech Group including evaluations for spoken dialog systems, automatic speech recognition, and machine translation. She is currently the Coordinator for the NIST Language Recognition Evaluation.