Evaluations of Automatic Speaker Classification Systems

Alvin F. Martin

National Institute of Standards and Technology
100 Bureau Drive Stop 8940
Gaithersburg, MD 20899-8940
alvin.martin@nist.gov

Abstract. The annual NIST Speaker Recognition Evaluations (SREs) from 1996 to 2006 have been internationally recognized as the leading source or performance evaluation of research systems in the speaker classification field. We discuss how these evaluations have developed and been conducted and the performance measures used. We consider the key factors that have been studied for their effect on performance, including training and test durations, channel variability, and speaker variability. We examine the extent to which progress has been observed in state-of-the-art performance. We also consider how the technology has changed over the past decade, other evaluations that have been conducted or planned, and where the field may be headed in the future.

Keywords: speaker recognition, speaker detection, speaker classification, speaker identification, speaker evaluation, NIST evaluations, NIST SRE, DET curves.

1 The Challenge

We consider the challenge of developing effective procedures for testing and evaluation of automatic speaker classification systems. This is a developing field of technology, and one with significant commercial potential. Such a field does not readily lend itself to objective technical evaluation, particularly in its early development.

Speaker recognition has developed somewhat in the shadow of the field of automatic speech recognition, where the objective is to transcribe the words (and perhaps understand their meaning as well) of a particular, or preferably of any, speaker. The development of evaluation in this area may be instructive.

In the 1970's and 1980's a number of speech recognition companies were offering products and anticipating a growing market for their offerings. And how good were their products. Each company recognized the need to quantify their performance and, invariably, each reported a correct word recognition rate in the range of 95–100 %. Yet potential users of the technology soon came to realize that in real world application scenarios of interest to them, they were likely to find far lower word recognition rates.

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 313–329, 2007.

Aside from telling outright lies about their performance, which may have occurred, each vendor would collect test data under ideal conditions for the speech recognition application of interest to them. And each would make very sure that a high recognition rate was achieved with this data; they couldn't hope to compete if they reported otherwise.

Potential users of the technology were in a difficult position. Each vendor claimed superior performance, and presumably had achieved it for its own proprietary data. But since the data was not shared, the performance of the different vendors' systems could not be meaningfully compared. Insightful users would recognize that with their own data and their own application scenarios they would not achieve the kind of results being reported, but until they acquired systems and used them in-house, they would not know which system was likely to be best for them, and how well it might do. This made it difficult to decide if the new technology would be cost effective compared with existing procedures or competing technologies.

George Doddington perhaps made the first efforts to test the performance of then existing speech recognizers on a common database [1]. He collected a database of spoken digits at Texas Instruments and invited vendors to supply a version of their systems to be used in in-house testing.

Soon after that, interest in such evaluation of speech recognition technology was taken up at the National Institute of Standards, which later became the National Institute of Standards and Technology (NIST), in Gaithersburg, Maryland. NIST has conducted a series of evaluations of speech recognition on different types of speech data, concentrating in recent years on broadcast news and conversational telephone speech. These evaluations have typically initially reported rather high word error rates, which have been reduced as a particular type of evaluation has been continued over several years. Indeed, when such error rates have been reduced below 10 % or so, NIST has shifted its evaluation focus to more difficult types of speech.

Speaker recognition lacked such independent evaluation into the 1990's. Each research site would choose its own data to use. This sometimes involved the use of proprietary corpora not available to other systems. But at least a few common speech corpora were becoming available, and a popular choice was the TIMIT Corpus [2]. This was a corpus of high quality phonetically transcribed speech including multiple sessions from a number of speakers (as needed for speaker recognition) that had been collected at Texas Instruments.

In 1994 the first of series of international workshop on speaker recognition was held in Martigny, Switzerland. It was followed by a similar workshop in Avignon, France in 1998. The third such workshop, in Crete in 2001 was dubbed "2001: A Speaker Odyssey". The subsequent workshops, in Toledo Spain in 2004 and San Juan, Puerto Rico in 2006 have continued the Speaker Odyssey name. The first two pre-Odyssey workshops, however, were dominated by researchers reporting results, generally very good results, on proprietary data sets or on the TIMIT data. This was viewed as frustrating by those who wanted to see meaningful performance comparisons on more real-world type data.

It was in this context that in 1996 NIST initiated its series of annual speaker recognition evaluations. These have concentrated on the use of conversational telephone data from corpora collected by the Linguistic Data Consortium (LDC) [20]. The central speaker detection task has remained the same throughout the evaluations. A system is given speech data (training data) known to be from a given target speaker, and given a separate test segment of speech data. It must then determine whether the test data was spoken by the target speaker. An evaluation test consists of a (long) sequence of trials of this type. For each trial, the given target speaker, defined by the training data, is the only speaker "known" to the system.

The NIST speaker recognition evaluations are described in greater detail in further sections of this chapter. Their history encapsulates the progress and problems encountered in this area over the past decade. They document the level of performance of state-of-the-art systems for speaker detection involving text independent conversational speech transmitted over public telephone channels and the degree of performance improvement over the period. But the evaluations have changed over the years, with the variety of test conditions increased, and the problems addressed sometimes made harder due to changes in general telephone technology and to greater interest in more challenging conditions as the technology has improved.

2 The NIST Evaluations

As noted, the basic task in all of the NIST speaker recognition evaluation has been speaker detection. This means that each test consists of a sequence of trials, where each trial is defined by a target speaker and a test segment of speech. The target speakers are defined by training data provided for each such speaker. This training data may consist of one or several speech segments guaranteed to contain speech of the speaker. The test segment contains unknown speech. The system must determine if in fact this speech was spoken by the target.

For each trial the system must supply both a hard decision ('T' or 'F') in answer to this question. In addition a likelihood score is required that quantifies the decision. Higher scores should indicate greater probability that the test speech is by the target.

Trials where the target is speaking, those for which the correct decision is 'T', are target trials. Trials where the target is not speaking are non-target (or impostor) trials. System errors in target trials are misses, while those in non-target trials are false alarms. Thus a system has two basic error rates, the percentage of target trials that are misses (miss rate) and the percentage of non-target trials that are false alarms (false alarm rate).

The basic error metric in the NIST evaluations has been a linear combination of these two rates that has been denoted C_{DET}. It is defined as

$$C_{DET} = Norm_{Fact} * ((C_{Miss} * P_{Miss|Target} * P_{Target}) + (C_{FA} * P_{FA|NonTarget})) \quad (1)$$

Table 1. The cost function that has served as the primary metric in the NIST evaluations is based on assigned relative costs for each miss and each false alarm and an assumed target richness chosen for possible applications of interest

Cost of a miss	$C_{Miss} = 10$
Cost of a false alarm	$C_{FA} = 1$
Probability of a target	$P_{Target} = 0.01$
Probability of a non-target	$P_{Non-Target} = 1 - P_{Target} = 0.99$
Normalization factor ($Norm_{Fact}$) is defined to make 1.0 the score of a knowledge-free system that always decides "False" It detection cost $C_{default} = 10 * 100\ \% * 0.01 + 1 * 0.99 = 0.1$ So $Norm_{Fact} = 10$	

C_{DET} can be viewed as a cost function based on assigned costs for misses and false alarms and an assumed target richness. But the assigned cost and assumed target richness are essentially arbitrarily chosen parameters. (Note that PTarget need not, and does not, correspond to the actual percentage of target trials in the evaluation test sets.) The values selected are believed to be reasonable ones for some applications of interest. The low target richness may be particularly applicable to text-independent applications. For some other applications a higher value may be appropriate, but so may a higher relative cost for false alarms, so these may cancel each other out to some extent.

There has, however, been recent work on developing a more application independent type of metric that allows after evaluation examination of performance for any specific parameters of interest. This requires that the confidence scores provided be actual probabilities, or better, actual log likelihood ratios. The metric Cllr, and the ways it may be utilized, are discussed in [3]. Such scores, and the use of this metric, was an option for participants in the 2006 evaluation and will probably receive attention in future evaluations.

3 Evaluation Parameters

Having defined the evaluation task, choices need to be made about the data to be collected and utilized. Evaluations are heavily dependent upon the collection of appropriate and sufficient data. Each evaluation test is defined by a sequence of trials, and time and cost for collection is likely to be the limiting factor determining the number of trials to be included.

The most basic evaluation parameters defining the trials are the duration of the training and test speech segments, and the timing of their collection. The training data for each target speaker may be collected in one or more different sessions. The amount of training data (duration of training speech) is typically the same or greater than the amount of test speech used in a given trial. (At

least for single session training, the training and test speech used in each trial may be viewed as playing symmetric roles.)

NIST has used a speech activity detector to determine the approximate durations of speech in training and test segments. In earlier evaluations considerable effort was made to be fairly precise about the speech durations in each trial. In later evaluations interest shifted in large part to using longer speech durations (in particular whole conversation sides) with less precision. Also in earlier evaluations the training and test segments consisted of concatenated segments of speech (as determined by the speech activity detector) with non-speech portions of the signal excised. In later evaluations continuous segments without excision were used, though estimates were still made of actual speech duration.

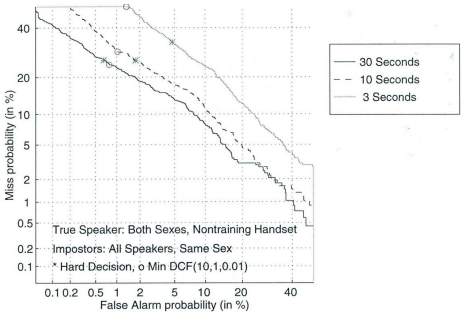

Fig. 1. Effect of test segment duration on performance, fixed durations

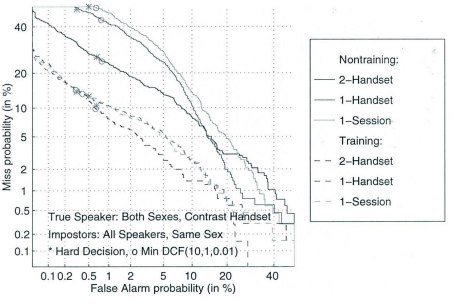

Fig. 2. Effect of match or non-match of training and test handsets, and of multiple training sessions with same or different handsets

Figures 1 and 3 show the effects of test segment duration on performance for a typical system in three different NIST evaluations. In all cases, we see the expected result of better performance with longer durations. In the early evaluations (Figure 1) the test segments had fixed approximate speech durations of 3, 10, or 30 seconds each. Later variable durations of up to a minute were used (Figure 3). Here it may be noted that the only strong effect on performance is seen for durations of less than 15 seconds.

With respect to training data, early NIST evaluations examined the effect of the number of training sessions, their diversity with respect to the telephone handsets used, and their relationship to the test segment handset for target trials. Figure 2 shows results for a system both where the test handset was the same as (one of) the training handsets and where it was not. (The duration of training speech is approximately the same for all six DET curves.) Most notable is the better performance when the same handset in used in training and test. (This is for target trials only; nontarget trials invariably involve different handsets.) Subsequent evaluations have emphasized different handsets, at least for landline transmission data. Examining the three curves where the test handset is different, it may be seen that having two training sessions yields better performance

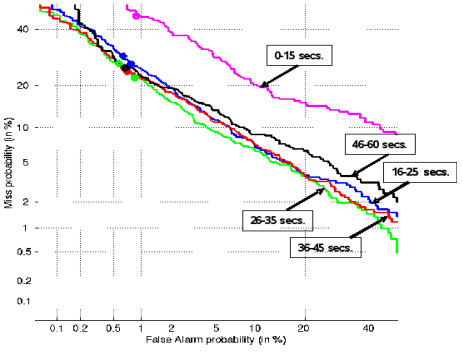

Fig. 3. Effect of test segment duration on performance, variable durations

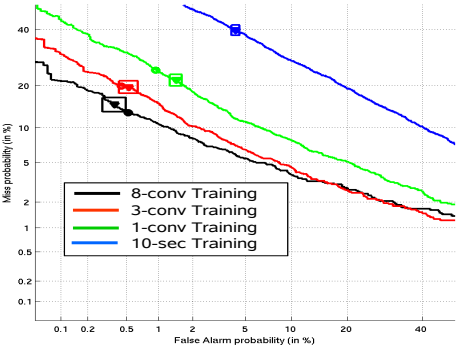

Fig. 4. Effects of varying amounts of training data on performance, all using the same handset

than one, and having these two sessions use different handsets further improves performance.

More recent evaluations have concentrated the effects of offering much larger amounts of training data. In figure 4 the curves show results when training consisted of 1, 3, or 8 whole conversation sides (each averaging about 2.5 minutes of speech). Also included is a 10-second training condition, which certainly remains of interest, particularly for some commercial applications. (In all cases the test segments consist of one conversation side of speech data.) The advantage of increased training data, where applications will support this is seen. It may also be seen that there is still a long way to go to achieve equivalent performance with very short segments of training data.

4 Channel Variability

Speaker recognition performance may be greatly enhanced by using a constant high-quality wideband channel, but the primary application interest of the technology is in its use over telephone channels, and perhaps over various types of differing and varying quality microphone channels. Thus the handling of channel variability is one of the key challenges to be overcome by the system designer and a key factor to be considered by the system evaluator.

The NIST evaluations, as noted previously, have until the last few years concentrated on telephone channels. But the nature of public telephone channels in the United States has changed considerably in recent years. The quality of traditional landline channels has improved. A decade or so ago carbon-button and electret microphones were both common in telephone handsets, and the early NIST evaluations considered the effects of handset microphone type on performance. Carbon-button microphones have become less common in recent years, but a bigger change has been the widespread use of cellular phones in the U.S. in recent years. Thus the recent evaluations have examined the performance effects of cellular as opposed to landline transmission.

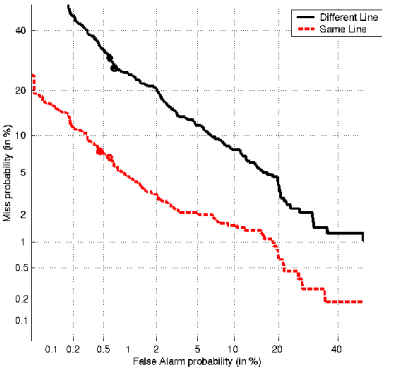

Fig. 5. Effect of same or different phone line (and presumably handset) in training and test

Figure 5, involving one system from an early evaluation, shows the effect of using a fixed or a variable telephone line, and presumably handset, in target trials. Clearly, having the same handset used in each speaker's training and test segments makes the problem far easier. But the use of caller id is simpler and more effective. (Note that non-target trials invariably involve the use of different phone lines and handsets unless special arrangements are made to do otherwise.) The situation of practical interest is where training and test phone lines differ, and later evaluations focused only on such cases, as least for landline trials.

Figure 6 shows the effects of microphone handset types for five different systems in an early evaluation. Two different effects are convolved to different overall effect in the different systems. In general performance is better with electret than with carbon-button handsets (the fourth system is something of an exception). But performance is also generally superior when the training and test handset types are the same. So the black curves generally show relatively good performance, and the red and blue curves relatively poor performance, while the green curves (all carbon-button) show variable performance.

Figure 7, from a recent evaluation, presents a similar type of plot for one system showing the effect of cellular or landline transmission in training and test. Perhaps not surprisingly, performance appears to be considerably better for landline data.

The most recent NIST evaluations have included some telephone conversations where the speech of one of the conversants was simultaneously recorded over a (cellular) telephone channel and over eight different microphone channels. Figure 8 shows performance results for one system involving the nine different representations of the same test conversations. (The training is fixed and recorded over a telephone channel.) The main point to be noted is that the telephone results are far superior to those of all the microphones. It should be noted that this was the first such NIST evaluation, and that cross-channel performance may be expected to improve in future evaluations.

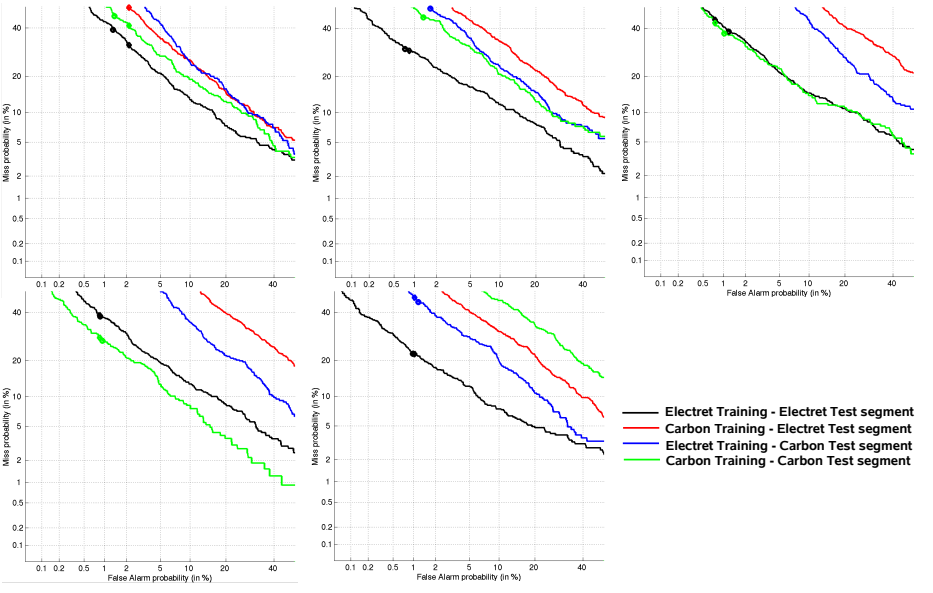

Fig. 6. Effect of using different combinations of handsets with carbon-button or electret microphones in training and test. These effects vary for the five different systems shown.

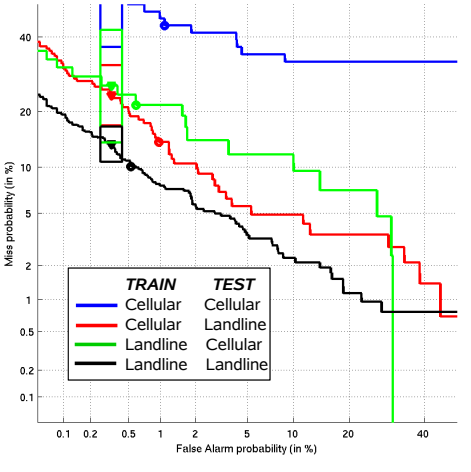

Fig. 7. Effects of using cellular or landline data in training and test on performance

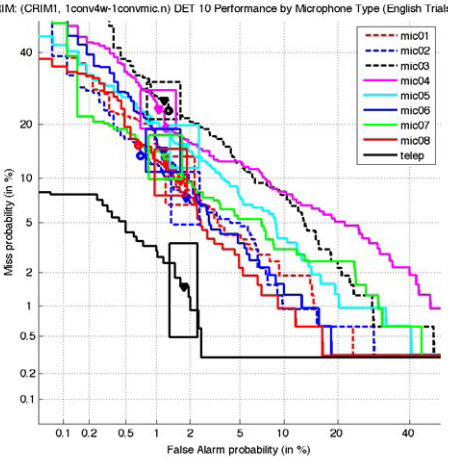

Fig. 8. Effect on performance of using any of eight different microphone channels or telephone data in the test segment, with training always on telephone data

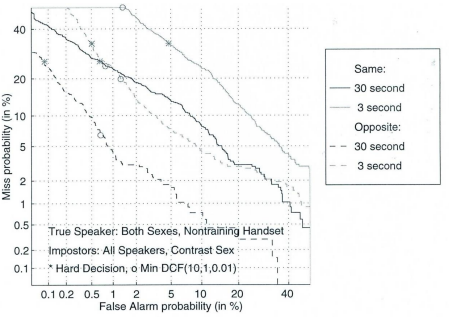

Fig. 9. Performance by whether non-target trials involve speakers of same or opposite sex for two durations

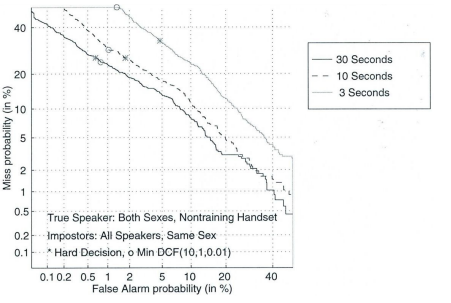

Fig. 10. Performance by language (English or non-English) in the training and test data for all trials

5 Speaker Variability

Variability between different speakers, a key problem for speaker independent word recognition, is the characteristic that makes speaker classification technology possible. A major division of speakers into two classes is by sex. Figure 9, from the first NIST evaluation, shows performance (for both 30-second and 3-second test segments) when the non-target trials involve speakers of same sex or of opposite sex. Since gender recognition tends to be highly accurate, the results are as might be expected. Including cross-sex trials in evaluations is one way to show better results. Subsequent NIST evaluations have excluded such trials.

The variability of individual speakers, on the other hand, is a major challenge to speaker classification technology. Speaker consistency is a highly desirable attribute for successful recognition, but in the real world speakers often do not maintain consistency for a variety of reasons. Voices change because of health problems (such as colds) and because of stress and emotional conditions. And in the long run they change as people age.

Measuring speaker variability in evaluation is not easy to do, as people cannot readily be instructed to demonstrate variability in their voices on demand. Creating stress conditions is not something that committees on the use of human subjects look fondly upon. And data collection sessions far enough apart in time to reveal the effects of aging are not readily arranged.

Figure 11 explores one way of examining the effect of speaker variation. For one system in a particular evaluation, we estimated the speaker's average pitch in the training and in the test data. The figure shows the large performance difference between the quarter of the target trials where the speaker was most consistent in average pitch between training and test and the quarter of the trials (perhaps involving one session with a cold) where the speaker had the greatest relative pitch differences.

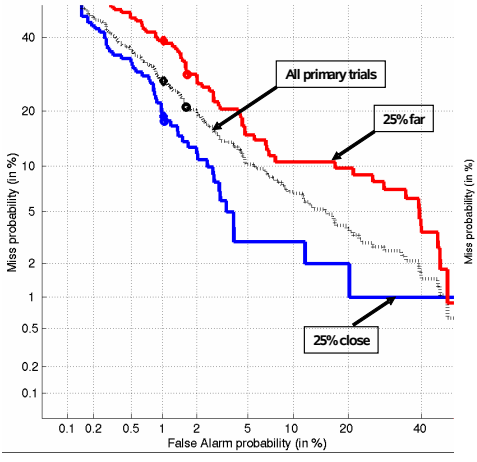

Fig. 11. Performance by relative closeness of training and test average pitch differences in target (same-speaker) trials

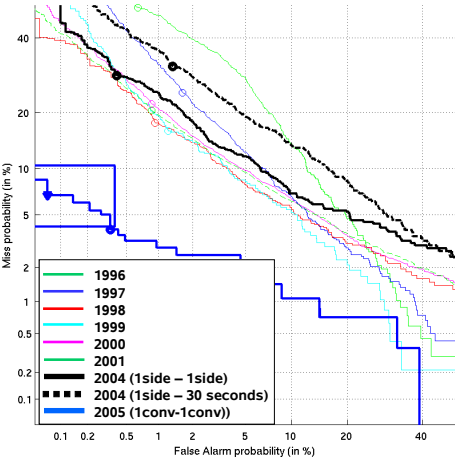

Fig. 12. Best-system performance history for landline trials 1996-2005. Prior to 2004 training generally consisted of two minutes of speech, and test of 30 seconds (an average of 30 seconds) of speech.

Another, more controllable way in which a speaker may vary, is in language. In recent NIST evaluations a number of bilingual speakers (of English and another language) were included. Figure 10 show performance results based on whether the training and the test speech were in English (E) or a non-English language (N). Clearly language consistency matters, at least for this system and others tested in this evaluation.

6 Measuring Progress

The primary purpose of evaluation of research systems in a developing field of technology such as speaker recognition is to encourage progress in the field. It is therefore of key concern to determine the degree of progress that has occurred over a period of years.

But there are difficulties in doing this. It can be hard to ensure that different test sets present equal task difficulty, even if they are chosen in substantially the same way. But evaluations do not remain constant from year to year. They change to reflect the changing interests and priorities of those who are sponsoring and organizing the evaluations. Improving system performance may be a reason to choose to make the task harder, thus appearing to suppress further performance improvement. And in the case of speaker recognition over telephone lines, changes in the public phone system affect the evaluation results. In particular, the increasing use of cellular telephones, which we have seen have an adverse effect on performance, has made comparisons more difficult.

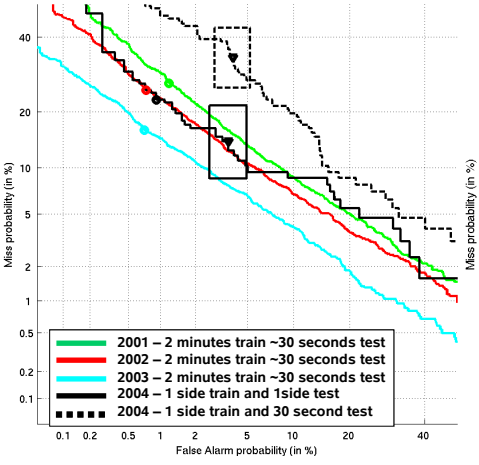

Fig. 13. Best-system performance history for cellular trials 2001-2004

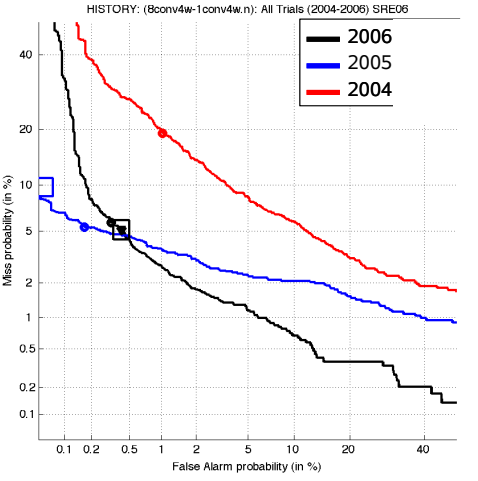

Fig. 14. Best-system performance history for eight conversation training (extended training data) 2004-2006

The NIST evaluations may be divided into three phases. From 1996 to 2001 the data was primarily landline, selected from the conversations of the several Switchboard corpora collected by the LDC. The primary test condition involved two minutes of training and thirty second test segments (variable averaging thirty seconds in 2001). For 2001 to 2003 testing similar but on the Switchboard cellular corpora (both landline and cellular were used in 2001). Since 2004 the LDC Mixer Corpus [16,17] has been used, with a different collection protocol, a mix of landline and cellular data, and some calls in languages other than English. Table 2 summarizes these three phases in the data used in the evaluations.

For each evaluation an effort is made to assess the overall level of performance improvement (or the lack thereof) between the best performing systems of the current and preceding years, matching test conditions of interest to the extent possible, and NIST has regularly sought to do this. Figures 12 – 15 attempt to suggest the degrees of progress that have been observed over the course of the NIST evaluations.

Figure 12 presents best system results on trials involving landline data between 1996 and 2005. (2002 and 2003 are omitted because the great majority of trials those years involved cellular data.) The results tend to divide between those for years prior to 2002 and those for years after 2003. For the earlier years, there was clear progress from 1996 to 1998, and then somewhat of a plateau until 2001. The Mixer data used starting in 2004 resulted in an apparent adverse performance effect, even with increased training and test durations. Two different test conditions in 2004 show better performance with longer duration test data, as expected. The number of all landline trials was limited in 2005, but a considerable performance improvement is observed.

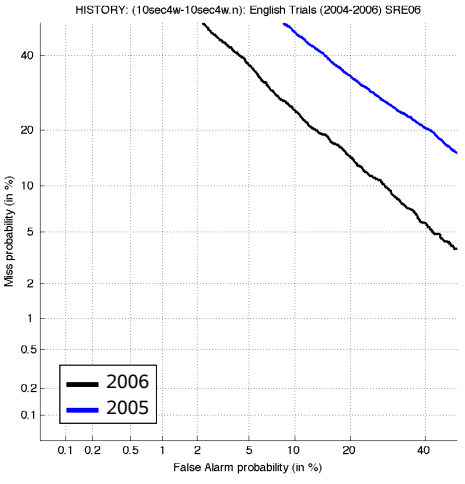

Fig. 15. Best-system performance history for short duration (10-second) training and test data trials 2005-2006

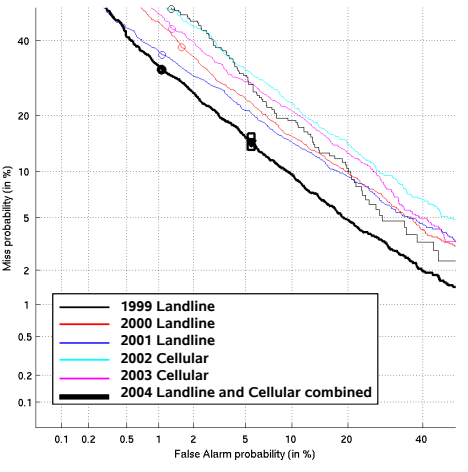

Fig. 16. Best-system performance history for two-speaker training and test trials 1999-2004

Figure 13 gives a similar history plot of best performing systems on cellular data from 2001 to 2004. There is general progress from 2001 to 2003. For 2004 there are two different test conditions, both of which are different from the conditions of the preceding years, and the number of cellular trials was smaller than before, making the curves less smooth. Moreover, 2004 was the first year in which Mixer data was used, adapting to which may have been a challenge for systems. In any case, the best 2004 performance did not match the best of 2003.

Since 2001, when George Doddington demonstrated the potential gains from exploiting high level idiolectal type information for speaker recognition from longer durations of speech [4,5], a major focus of the evaluations has been on the level of performance that may be achieved by the use of "extended duration" speech, particularly for training. Recent evaluations have included a condition on training on eight different conversation sides of each target (averaging about 2.5 minutes of speech each, while testing on single whole conversation sides. The previous discussion on duration has noted the effect of extended training on performance. Figure 14 shows results for the best performing system for the past three years. Results for earlier years are not comparable, because only with the Mixer data of these recent years was it possible to assure that the test handsets were distinct from those used in training. There was a considerable improvement in 2005 over 2004, and a more mixed result in 2006 compared with 2005. It is believed that the shape of the 2006 DET curve may be due to the presence of more trials involving non-English speech in 2006 than in 2005. This is another confounding factor in judging performance improvement.

Short duration training and test has been included in the NIST evaluations largely by popular demand. While performance is much inferior when training

Table 2. Corpora used and primary tests in three phases of the NIST SREs

1996-2001	Switchboard-1, Switchboard-2 Phases I, II, III	2-minute training (1 or 2 sessions), 3, 10, 30 second test segments, variable duration test segments in 2001 (averaging 30 sec.)
2001-2003	Switchboard Cellular Parts 1, 2	2-minute single session training, variable (15-45 sec.) duration test segments
2004-2006	Mixer (including some non-English conversations and multi-channel microphone data in 2005-2006	8, 3, or 1 conversation side training, 1 conversation side test segments (also 10 sec. training and test)

and test are limited to ten second speech durations, there is considerable commercial potential in being able to achieve good results in this case. Figure 15 shows that considerable improvement was seen in the best evaluation systems between 2005 and 2006, but that there remains a long way to go to achieve performance acceptable for most applications.

7 Multi-speaker

Speaker recognition in a multi-speaker environment, a subject perhaps outside the mainstream of work in speaker classification, has been a part of the recent NIST evaluations. They have focused on the summed channel situation where the input consists of the combined two channels of a phone conversation between two persons. The target speaker training data may be single channel, but the recent NIST evaluations have included a training condition consisting of three conversations involving the target speaker with three different people, requiring systems to find and segment the target speech in the training conversations.

Figure 16 shows a history plot of best systems for the two-speaker condition involving both landline and cellular data from 1999 to 2004. It shows a rather satisfying record of improvement for each type, with the best results occurring in 2004 on data involving both landline and cellular calls.

Earlier NIST evaluations also had tasks specifically for speaker segmentation and tracking within multiple speaker speech [23]. This kind of task has since been pursued in other in other evaluations, including the speaker diarization task of the NIST Rich Transcription Meeting Room evaluations [6,7,8] and the internationally (U.S. and Europe) based CLEAR (Classification of Events, Activities, and Relationships) [9] evaluations.

8 Other Evaluations

The author's perspective is oriented toward the NIST evaluations, and these have certainly assumed the leading role in the field to date, but there have been other evaluations, and there will undoubtedly be further ones.

In 2003 TNO, a Dutch applied scientific research organization, sponsored an evaluation of forensic speaker recognition. They were able to obtain, for limited evaluation use, appropriate audio data from actual police investigations. There were a variety of test conditions, involving different durations and types of data, and participant were asked for decisions in a sequence of trials using a format based on that used in the NIST evaluations. Some of its results are described in [10].

Another, if somewhat less successful evaluation, was held in conjunction with the Odyssey 2001 workshop in Crete. A couple of evaluation tracks were offered to participants in connection with the workshop. One involved a subset of the previous year's NIST evaluation. NIST analyzed submitted results much as in its regular evaluations. See [11,12]. The other track involved text-dependent speaker verification, where the enrollment and verification data consisted of speakers saying one of 17 specified passwords. This track is discussed in [13,14]. Participation was limited and, with respect to the second track involving spoken passwords, this perhaps may show the difficulty of creating text-dependent evaluations of general interest that can attract participants from commercial companies.

The use of speaker recognition as a biometric that may be used for secure verification of people's identities in light of recent word events is attracting increasing interest on both sides of the Atlantic. In Europe, however, there has been greater interest in using multiple biometrics, including speech, in combination to achieve increased performance. A major project denoted BioSecure, a part of the 6th Framework Programme of the European Community, is coordinating a multi-year interdisciplinary research program in support of this. It includes a "2007 BioSecure Evaluation Campaign" involving the use of voice, face, signature, fingerprint, hand, and iris data in a multi-faceted effort that is to launch in March, 2007 [15].

9 Future of Speaker Evaluation

After annual NIST evaluations from 1996 to 2006 it was decided, for a variety of reasons not to hold an evaluation in 2007. The evaluations have become larger over the years, both in test set size and number of participants, and more complicated in terms of the variety of tests included. The hiatus will provide additional time for data collection, always the key limiting factor in evaluation planning. This will allow the next evaluation to include considerably more data corresponding to cross-channel evaluation conditions. The hiatus is also intended to allow time to recruit an additional person to support the evaluation, but it remains to be seen whether continuing annual evaluations will be seen as feasible.

But speaker detection is an area of growing interest, and future evaluations, coordinated by NIST and perhaps other organizations appears quite certain.

There is a likelihood of growing government funding to support research in the area both in the United States and the European Union. This is expected to result in expanded evaluations in the United States while, as noted previously, there are plans in Europe for expanded evaluation of the fusion of biometric technologies including speaker.

The development of the technology may also produce increased demand for more product oriented evaluation. Very high performance, as noted, can be achieved for somewhat limited conditions, and systems to support these will become more visible in the commercial marketplace. But for the more challenging aspects of the task, with full text-independence and the use of the public telephone network or across multiple channels, there remain considerable performance limitations and a continuing need for ongoing evaluation of research systems.

References

1. Doddington, G.: Speech Recognition: Â turning theory to practice. IEEE Spectrum 18(9), 26–32 (1981)
2. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallet, D.S., Dahlgren, N.L.: The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. Technical report, National Institute of Standards and Technology, Gaithersburg (1993)
3. Brümmer, N., Du Preez, J.: Application-independent evaluation of speaker detection. Computer Speech & Language 20(2-3), 230–275 (2006)
4. Doddington, G.: Speaker recognition based on idiolectal differences between speakers. In: Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001). Aalborg, Denmark, Vol. 4, pp. 2521–2524 (2001)
5. Andrews, W.D., Kohler, M.A., Campbell, J.P., Godfrey, J.J.: Phonetic, idiolectal, and acoustic speaker recognition. In: Proceedings of the the Odyssey Speaker Recognition Workshop (Odyssey 2001), Chania, Crete, Greece, pp. 55–63 (2001)
6. Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D.: Sheep, Goats, Lambs and Woves: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation. In: Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP '98) (1998)
7. Fiscus, J.: The NIST Rich Transcription Evaluation Series, NIST web-site (2007), `http://nist.gov/speech/tests/rt/`
8. Fiscus, J., Radde, N., Garofolo, J.S., Le, A., Ajot, J., Laprun, C.: The Rich Transcription 2005 Spring Meeting Recognition Evaluation. In: Renals, S., Bengio, S. (eds.) MLMI 2005. LNCS, vol. 3869, Springer, Heidelberg (2006)
9. CLEAR2007: Classification of Events, Activities and Relationships, Evaluation and Workshop (2007), `http://www.clear-evaluation.org/`
10. Van Leeuwen, D.A., Martin, A.F., Przybocki, M.A., Boutenc, J.S.: NIST and NFI-TNO evaluations of automatic speaker recognition. Computer Speech & Language 20(2–3), 128–158 (2006)
11. Hansen, E.G., Slyh, R.E., Anderson, T.R.: Formant and F0 Features for Speaker Verification. In: Proceedings of the the Odyssey Speaker Recognition Workshop (Odyssey 2001), Chania, Crete, Greece, pp. 25–29 (2001)

12. Przybocki, M.A., Martin, A.F.: Odyssey Text Independent Evaluation Data. In: Proceedings of the the Odyssey Speaker Recognition Workshop (Odyssey 2001), Chania, Crete, Greece, pp. 21–23 (2001)
13. Higgins, A.L., Bahler, L.G.: ITT SpeakerKey Evaluation. In: Proceedings of the the Odyssey Speaker Recognition Workshop (Odyssey 2001), Chania, Crete, Greece, pp. 31–32 (2001)
14. Toledo-Ronen, O.: Speech Detection for Text-Dependent Speaker Verification. In: Proceedings of the the Odyssey Speaker Recognition Workshop (Odyssey 2001), Chania, Crete, Greece, pp. 33–36 (2001)
15. BioSecure: BioSecure Evaluation Campaign (2007), `http://www.biosecure.info/eval/`
16. Campbell, J.P., Nakasone, H., Cieri, C., Miller, D., Walker, K., Martin, A.F., Przybocki, M.A.: The MMSR Bilingual and Crosschannel Corpora for Speaker Recognition. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC '04), Lisbon, Portugal, [Alvin: wasn't this one published at the Odyssey 04 workshop raher than LREC?] (2004)
17. Cieri, C., Andrew, W., Campbell, J.P., Doddington, G., Godfrey, J., Huang, S., Libermann, M., Martin, A., Nakasone, H., Przybocki, M., Walter, K.: The Mixer and Transcript Reading Corpora: Resources for Multilingual Crosschannel Speaker Recognition Research. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06), Genoa, Italy (2006)
18. Reynolds, D.A., Doddington, G., Przybocki, M., Marin, A.: The NIST speaker recognition evaluation - overview, methodology, systems, results, perspectives. Speech Communication 31(2-3), 225–254 (2000)
19. Fiscus, J., Ajot, J., Michel, M., Garofolo, J.S.: The Rich Transcription 2006 Spring Meeting Recognition Evaluation. In: Renals, S., Bengio, S., Fiscus, J.G. (eds.) MLMI 2006. LNCS, vol. 4299, Springer, Heidelberg (2006)
20. Linguistic Data Consortium: Catalog of Speaker Recognition Corpora (2007), `http://www.ldc.upenn.edu/Catalog/`
21. Martin, A.F., Przybocki, M.A.: The NIST 1999 Speaker Recognition Evaluation - An Overview. Digital Signal Processing 10, 1–18 (2000)
22. Martin, A.F., Przybocki, M.A.: The NIST Speaker Recognition Evaluations: 1996-2001. In: Proceedings of the the Odyssey Speaker Recognition Workshop (Odyssey 2001), Chania, Crete, Greece, pp. 39–43 (2001)
23. Martin, A.F., Przybocki, M.A., Doddington, G.: Speaker Recognition in a Multi-Speaker Environment. In: Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001), Aalborg, Denmark, vol. 2, pp. 787–790 (2001)
24. Martin, A., Miller, D., Przybocki, M., Campbell, J.: Conversational Telephone Speech Corpus Collection for the NIST Speaker Recognition Evaluation 2004. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC '04), Lisbon, Portugal (2004)
25. Martin, A.F., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M.: The DET Curve in Assessment of Detection Task Performance. In: Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech 1997). Rhodes, Greece, vol. 4, pp. 1985–1988 (1997)
26. Martin, A.F., Przybocki, M.A., Campbell, J.P.: The NIST speaker recognition evaluation program. In: Wayman, J., Jain, A.K., Wayman, D.M. (eds.) Biometric Systems: Technology, Design and Performance Evaluation, pp. 241–262. Springer, Heidelberg (2005)

27. Martin, A.F., Przybocki, M.A., Le, A.N.: The NIST Speaker Recognition Evaluation Series, NIST web-site (2007), http://www.nist.gov/speech/tests/spk/
28. Philipps, P.J., Martin, A., Wilson, C., Przybocki, M.: An introduction to evaluating biometric systems. IEEE Computer 33(2), 56–63 (2000)
29. Przybocki, M.A., Martin, A.F.: NIST speaker recognition evaluation. In: Proceedings of the Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C '98), Avignon, pp. 120–123 (1998)
30. Przybocki, M.A., Martin, A.F.: NIST Speaker Recognition Evaluation Chronicles. In: Proceedings of the ODYSSEY Speaker and Language Recognition Workshop (Odyssey 04), Toledo, Spain (2004)
31. Przybocki, M.A., Martin, A.F.: NIST's Assessment of Text Independent Speaker Recognition Performance. In: The Advent of Biometrics on the Internet: Proceedings of the COST 275 Workshop, Rome, Italy, pp. 25–32 (2000)
32. Przybocki, M.A., Martin, A.F., Le, A.N.: NIST Speaker Recognition Evaluation Chronicles Part 2. In: Proceedings of the ODYSSEY Speaker and Language Recognition Workshop (Odyssey '06), San Juan, Puerto Rico (2006)