NIST Speaker Recognition Evaluation Chronicles – Part 2

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

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

Abstract

NIST has coordinated annual evaluations of text-independent speaker recognition since 1996. This update to an Odyssey 2004 paper concentrates on the past two years of the NIST evaluations. We discuss in particular the results of the 2004 and 2005 evaluations, and how they compare to earlier evaluation results. We also discuss the preparation and planning for the 2006 evaluation, which concludes with the evaluation workshop in San Juan, Puerto Rico, in June 2006.

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] [2]. The evaluations have been posed primarily as detection tasks relying on various conversational telephone speech corpora as the main source of evaluation data.

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 remained the primary evaluation focus.

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 [3]), 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

An evaluation test consists of a series of *trials* in each of which the system must determine whether a given *model speaker*, defined by specified training speech data, is speaking in a given test segment of speech. Test trials can be categorized as either *target trials*, meaning the target speaker is speaking in the test segment, or *impostor* (non-target) *trials*, meaning the target 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. 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:

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

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 equation 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.

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 other operating points, as determined by the parameters of equation 1, where false alarm rates are much lower than miss rate. Note also that the cost function depends on the system's calibration of the tradeoff between misses and false alarms (the likelihood threshold), while a measure such as equal error rate inherently assumes an optimal calibration.

2.3. An alternative cost function

The ordering of the confidence scores is all that matters for computing the detection cost function as defined above and which, as noted, is application specific. But confidence scores can be more informative, and used to serve any application, if they represent actual probability estimates. For the 2006 evaluation [4], NIST has invited participants to provide as scores likelihood ratio values independent of the application parameters. In terms of the conditional probabilities for the observed data of a given trial relative to the alternative target and non-target hypotheses the likelihood ratio (LR) is defined by:

$$LR = \text{prob (data|target hyp.)} / \text{prob (data|non-target hyp.)} \quad (2)$$

Sites may, optionally, indicate that their scores should be interpreted as likelihood ratios.

1-4244-0472-X/06/$20.00 ©

A further type of scoring will be performed on submissions whose scores are declared to represent likelihood ratios. A log likelihood ratio (*llr*) based cost function, which is not application specific and may be given an information theoretic interpretation, is defined as follows:

$$C_{llr} = 1/(2*\log 2)*((\sum \log(1+1/s)/N_{TT}) + (\sum \log(1+s)/N_{NT})) \quad (3)$$

where the first summation is over all target trials, the second is over all non-target trials, N_{TT} and N_{NT} are the total numbers of target and non-target trials, respectively, and s represents a trial's likelihood ratio score. The reasons for choosing this cost function, and its possible interpretations, are described in detail in [5].

2.4. 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 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, suggested by Swets [6], 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 [7] in the 1996 evaluation [8], 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. 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.

For systems with likelihood ratio scores, NIST in 2006 will experiment with graphs, somewhat analogous to DET curves, based on the likelihood ratio cost function. These may serve to indicate the ranges of possible applications for which a system is or is not well calibrated. Again, see [5].

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 [9]. The several collections of Switchboard style corpora, each of which included hundreds of speakers and thousands of conversations, were extensively used in the detection tasks of the NIST Speaker Recognition Evaluations from 1996 to 2003.

The 2004 and 2005 evaluations, and the upcoming 2006 evaluation, have utilized more recent LDC collections of conversational telephone speech data based on the "Fishboard" platform [10] [11] [12].

4. Evaluation Tasks

The history of how the evaluation tasks and performance varied from 1996 to 2003 is reviewed in [1]; here we review tasks and performance in the 2004 and 2005 evaluations, and discuss how these results may be compared with performance in the earlier years of the evaluation.

Recent evaluations have concentrated on using whole conversation sides (averaging about two and half minutes per conversant) as test segments, with training on one or more such sides of a target speaker. In these recent evaluations NIST has also provided errorful ASR (automatic speech recognition) transcripts of all of the evaluation speech data. In 2005 the transcripts were estimated to have an error rate of around 20%.

We describe here results with one side of training (limited training) and results with eight sides of training (extended training). We also consider results where the data does not consist of single conversation sides, but rather of summed channel data with both conversation sides and both speakers present in the signal (two-speaker detection).

It should be noted that it has become the accepted community practice not to publicize evaluation 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. Therefore, the DET curve plots presented here show the best performing systems in different years for different evaluation conditions without identifying the sites that developed these systems.

4.1. One-Speaker Detection with Limited Training

This is considered the basic evaluation task, the one required of all participants. Both training and test consist of one conversation side of data, and the system must determine whether the test conversation segment contains the target speaker defined by the training.

There was a key change in evaluation protocol in 2005, compared with 2004 and previous years, however. The data supplied consists of single channel conversation sides, each with one speaker. In 2005, however, the other conversation side was also supplied, for both training and test data. Note that this other side data involved, in almost all cases, speech from two other speakers than those in the designated training and test conversation sides of interest. As in previous years, ASR transcripts of the supplied data were also made available. Thus the additional data could assist in modeling the nature of the conversations taking place. Most participating sites did not make use of this additional data, but at least a couple of sites did seek to use it to improve performance.

Figure 1 compares the best system performance in 2004 with the performance of seven better performing 2005 systems. It was quite an achievement that seven systems performed as well or better in 2005 as the best performing 2004 system. Note that on the normal deviate scale of the DET plots shown, the improvement is not at all small. Comparing the best 2005 system with the best 2004 system, the equal error rate was cut by about 50%, and at a 10% miss probability the false alarm rate was reduced by a factor of approximately fifteen.

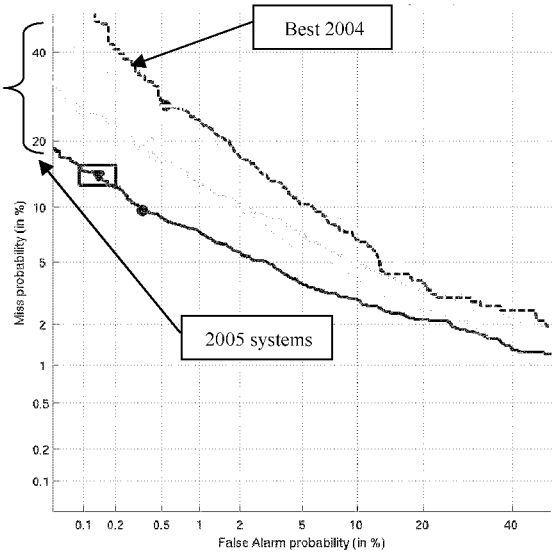

Figure 1: 2005 systems performing as well or better than the best 2004 system on one speaker detection with limited training.

We have noted that the provision of both channels of conversational data benefited certain systems, but most of the 2005 systems shown did not take advantage of this. It could be asked then whether the 2005 data was easier than the 2004 data, even though both data sets were selected from the same Mixer Corpus. A couple of sites ran their "mothballed" 2004 systems on the 2005 data. Figure 2 shows the performance comparison thus produced for one site. These results suggested that the 2005 data may have been slightly easier, but that this accounted for only a small part of the difference in the best systems' performance between 2004 and 2005.

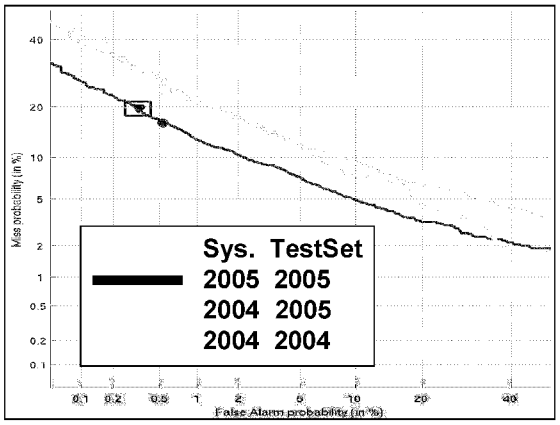

Figure 2: One site's mothballed system performance on the 2004 and 2005 evaluation data, along with its 2005 system performance

How do the 2004 and 2005 results compare with those of earlier evaluations? The earlier evaluations used conversational data from different corpora, including some involving all landline data and some involving primarily cellular data. Earlier evaluations generally used test segments with a maximum of 30 seconds of speech. And only during the past four years have ASR transcripts of data been made available, though sites have always been free to run ASR systems of their own. Discussion of one-speaker detection performance in the earlier evaluations may be found in both [13] and [1].

In Figure 33 we attempt to compare best system performance in 2004 and 2005 on landline data with best system performance on landline data in earlier evaluations. In 2005 the numbers of sides specifically known to have been recorded on landline data was limited, so this curve is less smooth than others, but the confidence box (of 95% limits) suggests that significantly better performance was achieved. Landline data was used with 30 second test segments in evaluations between 1996 and 2001. Landline data was not used in the primary evaluation condition in 2002 and 2003. For 2004 results were obtained with both whole conversation sides (about two and a half minutes of speech) and with 30 seconds of speech as test segments. Including both curves for 2004 helps to relate the earlier results to those of 2005.

The curves in Figure 33 show small differences over the course of 1998-2001. The poorer results for 30 second speech durations in 2004 than in several preceding years suggest that the task became harder with the Mixer data, which was the general view of participants. The advantage of whole conversation sides over 20-second segments in 2004 is apparent, but this difference is small compared to the whole conversation side improvement seen in 2005. A comparison of the best 2005 results in Figure 1 and Figure 33 gives a sense of how much the task eases by restricting to landline data only.

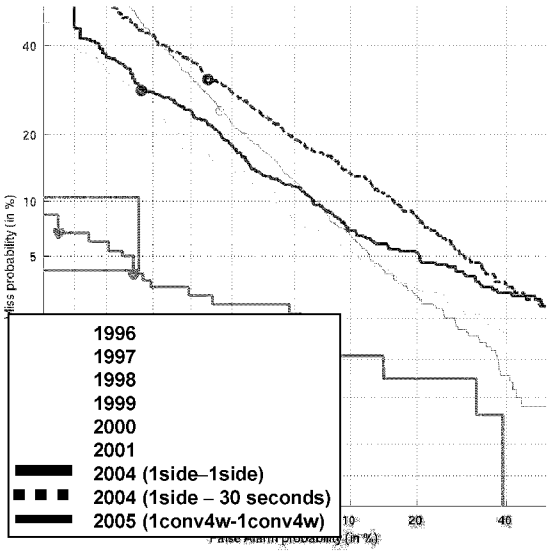

Figure 3: Best system performance on landline data in 2004 and 2005 as well as earlier years.

4.2. One-Speaker detection with extended training

For the past several years the NIST evaluation has included an extended data component, in part inspired by results reported in [14] and [15]. We have concentrated on a test

where eight training conversation sides are provided for each target speaker, along with a single conversation side as test segment. Figure 44 shows best system performance over the past four years. There was real performance improvement in 2005 over 2004, particularly in the low false alarm region of the DET plots, which is the region of greatest interest.

It may be seen that the best performance DET curve deteriorated in 2004 compared with the two preceding years. Different corpora from Mixer were used in the earlier years, and we believe that the Mixer data was more difficult in three specific respects. The Mixer target trials all involved a different phone number, and presumably a different telephone handset, in the training and the test data; in the Switchboard data used earlier this was not always the case. The Mixer training data involved a single phone number (handset) in all the training conversations, while the earlier data often had multiple training handsets. More training handset variation and, especially, lack of handset variation between training and test has previously been seen to aid performance. Finally, the Mixer data was a mixture of landline and cellular data, while the earlier data was all landline, which has been seen to make the task easier.

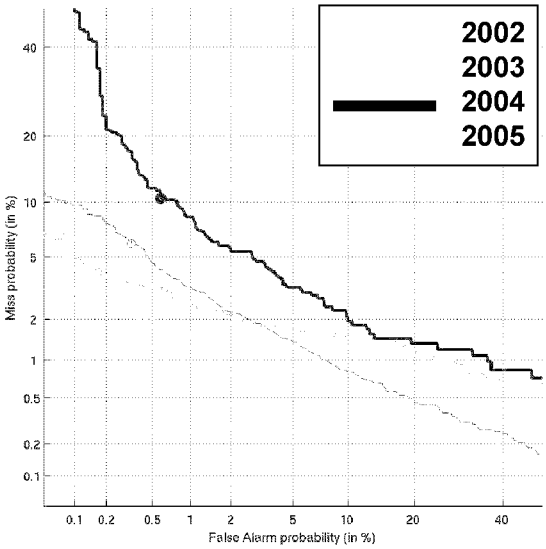

Figure 4: Best system performance with extended training 2002—2005.

4.3. Two-Speaker Detection

For some years the NIST evaluations have also included a two-speaker detection condition. Here the data, training and test, is summed two channel conversational data, 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.

Figure 5 shows the best performance results on the two-speaker test for each year from 1999 to 2005. General progress is apparent, with a setback in performance in 2002 and 2003 when the data switched to cellular from landline. Also, in the three earlier years, only the test segments consisted of summed channel data; the training was single channel. The best results were in 2004 and 2005 using a mixture of landline and cellular data, with a significant improvement in 2005 over 2004. Comparing Figure 55 with Figure 1 shows that there is still a gap between one-speaker and two-speaker performance, but it is not a wide as it was in earlier years.

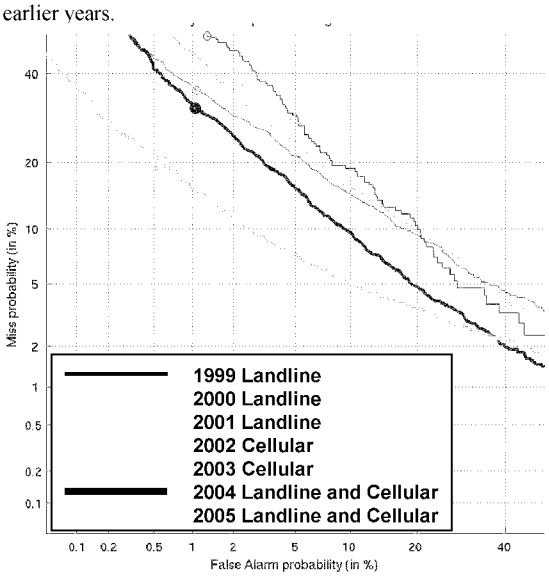

Figure 5: Best system performance for two speaker detection 1999—2005.2006 evaluation

5. 2006 Evaluation

The 2006 NIST Speaker Recognition evaluation was scheduled to be held in April and May of 2006. Each participating site was to receive the evaluation data, process it, and submit its results to NIST for scoring.

5.1. Evaluation data

The 2006 evaluation used a combination of newly collected conversational data and data recycled from the 2005 evaluation. The newly collected data was collected following the Mixer Corpus protocol also used in 2004 and 2005. A significant fraction of the speakers in the newly collected data were bilingual speakers who were asked to speak in their non-English language of fluency whenever they could be paired with speakers of the same language. This new data was supplemented with reused data from 2005 in order to increase the total number of English speakers in the test with at least eight training conversations. Also included was some previously collected but unexposed conversational data recorded simultaneously over both telephone and

microphone channels (described below) similar to the data of this type used in 2005.

5.2. Test conditions

The test conditions included in the 2006 evaluation were similar to those of recent years. There were five training and four test segment conditions, summarized in Table 1 and Table 2. All training data for each target speaker was selected to come from a different phone number, and presumably a different telephone handset, than did all test segment data from the speaker.

Note that one of the test segment conditions involves microphone recordings of conversational sides. As part of the Mixer collection the LDC has recorded some conversation sides simultaneously over eight microphone channels as well as over a telephone channel, using a custom-designed setup at several collection sites. It is hoped that in 2006 more participants will choose to do tests involving this condition, allowing study of the effects of cross-channel data on speaker recognition performance.

Table 1: Training conditions defined for the 2006 NIST evaluation.

Training Condition	Description
8 sides	8 conversation sides.
3 sides	3 conversation sides, generally subsets of 8-sides models.
1 side	1 conversation side.
10 seconds	A variable length segment containing about 10 seconds of speech. Each segment is taken from a 1 conversation side model.
3 conversations	3 summed-channel conversations. In general, the conversations include the sides of a 3-sides model.

Table 2: Test segment conditions defined for the 2006 NIST evaluation.

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

For 2006 it was decided to pare the matrix of tests included in the evaluation. Instead of the full matrix of 20 tests (5 training condition *x* 4 test segment conditions) as in 2005, the number of tests included was 15, as indicated by the matrix in Table 3. The tests not included were ones that had attracted few participants previously. Participating systems could do as many or few of these 15 as they 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 also offered 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). Unsupervised adaptation was an available option for each of the 15 tests. Results without adaptation were also required, permitting analysis of the performance effects of such adaptation.

Unsupervised adaptation was first introduced in the 2004 evaluation. Results were unimpressive that year, but one site obtained promising results in 2005. It is hoped that more participating sites will attempt it this year.

Table 3: Matrix of training and test segment conditions. The shaded entry is the required core test condition.

		Test Segment Condition			
		10 sec 2-chan	**1 conv 2-chan**	**1 conv summed -chan**	**1 conv aux mic**
Training Condition	**10 seconds 2-channel**	optional			
	1 conversation 2-channel	optional	required	optional	optional
	3 conversation 2-channel	optional	optional	optional	optional
	8 conversation 2-channel	optional	optional	optional	optional
	3 conversation summed-channel		optional	optional	

5.3. Participants

There were 39 organizations or teams of organizations registered to participate in the 2006 evaluation. This is by far a record number of participants for a NIST speaker recognition evaluation as in 2005, the previous record year, there were 27 participants. This presumably reflects the growing interest in this technology worldwide. The participating sites included research labs from companies,

non-profit organizations, governments, and universities in North America, Europe, the Middle East, Africa, Asia, and Australia.

5.4. Results

Full evaluation results will be presented at the NIST Evaluation Workshop in San Juan, Puerto Rico, in June 2006. The record number of participants suggests that this workshop will be very busy and crowded. Perhaps some very different and interesting new system approaches will be attempted in the evaluation and discussed at the workshop.

Summary results of the best performance achieved in the evaluation will be presented at the main Odyssey Speaker and Language Recognition Workshop immediately following the NIST workshop. This presentation will also include analysis of the effects on performance of various factors including training and test segment duration, language, telephone transmission type, and handset type. The results of the 2006 evaluation will also be compared with those of 2005 and earlier years.

6. Future evaluations

The NIST evaluations are expected to continue in future years. The success of such evaluations depends critically, as always, on collecting appropriate and sufficient data. The current collection paradigm appears well adapted for future plans, but cost is always an issue.

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] Przybocki, M. A. and Martin, A. F., "NIST Speaker Recognition Evaluation Chronicles", *Proc. Odyssey '04*, Toledo, Spain, June 2004.

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

[3] Martin, A. F., et al., "NIST Language Technology Evaluation Cookbook", *Proc. LREC '04*, Lisbon, Portugal, May-June 2004

[4] "The NIST Year2006 Speaker Recognition Recognition Evaluation Plan", *http://www.nist.gov/speech/tests/spk/2006/sre-06_evalpl a n -v1.pdf*, December 2005.

[5] Brummer, N. and du Preez, J., "Application-Independent Evaluation of Speaker Detection" in *Computer Speech & Language*, volume 20, issues 2-3, April-July 2006, pp. 230-275.

[6] Swets, J., ed., *Signal Detection and Recognition by Human Observers*, John Wiley & Sons, Inc., 1964, pp. 611-648.

[7] Martin, A. F., et al., "The DET Curve in Assessment of Detection Task Performance", *Proc. Eurospeech '97,* Rhodes, Greece, September 1997, Vol. 4, pp. 1899-1903.

[8] Martin, A. F., 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.*

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

[10] Campbell, J., et al., "The MMSR Bilingual and Crosschannel Corpora for Speaker Recognition Research and Evaluation", *Proc. Odyssey '04*, Toledo, Spain, June 2004

[11] Martin, A. F., et al., "Conversational Telephone Speech Corpus Collection for the NIST Speaker Recognition Evaluation 2004". *Proc LREC 2004*, Lisbon, Portugal, May-June 2004

[12] Cieri, C., et al., "The Mixer and Transcript Reading Corpora: Resources for Multilingual Crosschannel Speaker Recognition Research", Language Resources and Evaluation Conference (LREC), Genoa, Italy, May 2006

[13] Przybocki, M. A., and Martin, A. F., "NIST's Assessment of Text Independent Speaker Recognition Performance", *Proc. COST 275 Workshop – The Advent of Biometrics on the Internet*, Rome, Italy, November 2002, pp. 25-32

[14] Doddington, G., "Speaker Recognition Based on Idiolectal Differences Between Speakers", *Proc. Eurospeech '01,* Aalborg, Denmark, September 2001, Vol. 4, pp. 2521-2524

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