# Design Issues in Text-Independent Speaker Recognition Evaluation

# Alvin Martin, Mark Przybocki

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

#### Abstract

We discuss various considerations that have been involved in designing the past five annual NIST speaker recognition evaluations. These text-independent evaluations using conversational telephone speech have attracted state-of-the-art automatic systems from research sites around the world. The availability of appropriate data for sufficiently large test sets has been one key design consideration. There have also been variations in the specific task definitions, the amount and type of training data provided, and the durations of the test segments. The microphone types of the handsets used, as well as the match or mismatch of training and test handsets, have been found to be important considerations that greatly affect system performance.

#### 1. Introduction

During the past five years NIST (the National Institute of Standards and Technology) has coordinated evaluations of text-independent automatic speaker conversational telephone recognition using (Doddington et al., 2000; Martin and Przybocki, 2000; Przybocki and Martin, 1999). These evaluations have had the objectives of advancing the core technology and measuring current state-of-the-art performance. have therefore addressed the basic decision tasks in the field that are generally beyond the capabilities of existing commercial systems. The requirement of textindependent performance avoids association with any particular specific applications and enables the use of corpora containing realistic conversational or spontaneous speech.

NIST has created detailed written evaluation plans for each evaluation. These may be found, along with some information on performance results, on the speaker evaluation web site (NIST, 2000).

The following sections discuss some specific evaluation issues and constraints that have motivated the design of these evaluations. Obtaining sufficient and appropriate data for meaningful evaluation has been one major concern. Over the years there have been additions to the basic task definitions, as well as variations in the

amount of training data provided and the test segment durations. There have also been changes in the emphasis given to the properties of the telephone handsets used, which have been found to be important factors in system performance.

# 2. Data Availability

The evaluations have used the Switchboard and Switchboard-2 conversational telephone corpora. These corpora were collected by the Linguistic Data Consortium (LDC, 2000) with the needs of speaker recognition evaluation, as well as speech recognition evaluation, in mind. They were designed to include multiple conversations by most participating speakers, and the use of multiple telephone handsets by many speakers. Some properties of these corpora are summarized in Table 1.

The 2000 NIST evaluation will reuse much of the Switchboard-2 data used in the 1997 and 1998 evaluations. Reusing data clearly has disadvantages, but appropriate available data is limited and costly to produce. This recycling of data will provide a large and varied evaluation data set, with the 1999 data available for system development.

The 2000 evaluation will also include the Spanish language Ahumada corpus (Ortega-Garcia, 2000). This corpus features about 100 male speakers and multiple telephone sessions that include spontaneous, though not

	Switchboard-1	Swbd-2 Phase-1	Swbd-2 Phase-2	Swbd-2 Phase-3
Conversations	2866	3702	4575	2728
Encoding	2-channel mu-law	2-channel mu-law	2-channel mu-law	2-channel mu-law
Speakers	543	661	627	640
Talk on topic?	Yes	Frequently not	Frequently not	Frequently not
Use different phone numbers (handsets)?	Some	All initiated calls	All initiated calls	All initiated calls
Predominant age range	Varied, 18 and older	College age	College age	College age
Predominant Location	Varied over U.S.	Northeastern U.S.	Midwestern U.S.	Southern U.S.
Year used	1996	1997, 2000	1998, 2000	1999

conversational, speech.

NIST hopes to include cellular telephone speech data in future evaluations. The LDC is now collecting a Switchboard style cellular corpus. We would also like to include data in a variety of languages to the extent that suitable corpora, with sufficient conversational telephone speech from many speakers can be obtained.

#### 3. Tasks to Perform

NIST has tried to concentrate on what it regards as the basic decision tasks of speaker recognition not specific to any particular application. Each evaluation has included the basic task of *one-speaker detection*: given a segment of speech by a single speaker and a hypothesized known speaker (for whom training data is separately provided), determine whether or not the hypothesized speaker is the speaker in the segment.

In one-speaker trials, NIST always opts to have the segment speaker and the hypothesized speaker be of the same sex. Not including cross sex trials gives a more realistic view of system performance capabilities.

In 1999 two additional tasks were added. The *two-speaker detection* task uses channel summed segments of conversational speech involving two speakers. Systems are asked to determine whether a single specific speaker is present in the segment. The *speaker tracking* task takes similar conversational segments and asks systems to determine those time intervals within it, if any, where a specific hypothesized speaker is speaking.

In 2000 a fourth task, denoted *speaker segmentation*, is being added, which may be regarded as a variant of the tracking task. Here systems must divide conversational telephone speech segments involving multiple (possibly more than two) unknown speakers into speech intervals labeled by distinct speakers. This is not a decision task, but it forms a front-end to most approaches to the tracking task.

Table 2 summarizes these tasks.

#### 4. Evaluation Metric

A decision task necessarily involves two types of trials and two types of error:

- Target trials hypothesized speaker is present in test segment - errors are missed detections
- Impostor trials hypothesized speaker is not present in test segment - errors are false alarms

An overall metric must balance these two types of error. NIST has chosen to use a *detection cost function* of the form:

$$C_{Det} = (C_{Miss} * P_{Miss|Target} * P_{Target}) +$$

$$(C_{FalseAlarm} * P_{FalseAlarm|NonTarget} * P_{NonTarget})$$

Here  $P_{Miss|Target}$  is the missed detection rate and  $P_{FalseAlarm|NonTarget}$  is the false alarm rate,  $C_{Miss}$  is the cost of a missed detection,  $C_{FalseAlarm}$  is the cost of a false alarm,  $P_{Target}$  is the a priori probability of a target speaker, and  $P_{NonTarget}$  is the a priori probability of a non-target speaker. This formulation allows the evaluation to contain a different mix of target and impostor trials than the mix imagined for likely applications and for which the systems are asked to be optimized.

Along with this official metric, systems are also required to gives numerical likelihood scores to each decision. NIST uses these to present most results in the form of DET curves (Martin et al., 1997) showing all possible operating points based on all possible likelihood scores decision thresholds. Numerous examples may be seen on the NIST speaker recognition evaluation web site (NIST, 2000).

### 5. Test Set Size

A key aim of the NIST evaluations has been to obtain test set sizes large enough to give statistically meaningful results. The Switchboard Corpora are distinctive in offering thousands of conversations involving hundreds of speakers making large evaluation data sets possible. The numbers of trials needed for statistical significance is discussed in (Doddington, 1998). In particular this paper presents, based on the binomial distribution and an assumption of independent trials, "the rule of 30":

To be 90 percent confident that the true error rate is within  $\pm$ -30 percent of the observed error rate, there must be at least 30 errors.

This assumes, as also noted by Doddington, that target trials are independent of one another, an assumption somewhat compromised by the use of an average of about ten speech segments from each speaker (all from different conversations).

Based on these considerations, the NIST evaluations have perhaps been large enough overall, but when restricted to primary conditions of particular interest

Name	Test Segment	Task	Years
One-speaker detection	Concatenated single channel segments of one speaker	Determine whether hypothesized speaker is speaking	1996- 2000
Two-speaker detection	One minute of summed two- channel conversation involving two speakers	Determine whether hypothesized speaker is one of the speakers	1999- 2000
Speaker tracking	Same as two-speaker detection	Determine intervals (if any) where hypothesized speaker is speaking	1999- 2000
Speaker Segmentation	Same as two-speaker detection, but duration may exceed one minute	Determine the number of speakers and the intervals where each is speaking	2000

Table 2: Evaluation Tasks

Evaluation	Speakers	Target Trials	Impostor Trials
1996	40	1,427	98,172
1997	401	4,499	50,004
1998	497	5,000	45,000
1999 one-speaker detection	539	3,157	34,463
1999 two-speaker detection	539	3,158	34,748
1999 speaker tracking	507	1,998	2,002
2000 one-speaker detection	851	5,768	54,688
2000 two-speaker detection	851	5,768	54,688
2000 speaker tracking	777	2,000	2,000

Table 3: Evaluation Test Set Sizes

they have not been as large as might be desired. The 2000 evaluation, by reusing much of the previously used Switchboard-2 data, will have notably larger numbers of speakers and target trials. Moreover, all of the target trials will involve different training and test handsets (see section 8) and most will involve electret type handset microphones (see section 9). These will be conditions of primary interest for this evaluation. Table 3 shows the test set sizes of recent evaluations.

# 6. Training conditions

The amount and variability of training data is likely to affect performance. In recent evaluations NIST has provided two minutes or more of training data for each speaker. More important than the amount of data, it appears, is whether the data is all from one session, or from two different sessions, and whether the different sessions use the same or different handsets. Table 4 summarizes the training conditions included in recent evaluations

Since the effects of the training conditions were clearly demonstrated in previous evaluations, NIST decided in the 1999 and 2000 evaluations to simplify matters by using only one, single training condition, namely two-session training in 1999 and one-session training in 2000. Multiple-session training, it should be noted, may be unacceptable to users in many typical applications.

# 7. Test Segment Durations

Test segment duration is clearly a factor of interest in evaluating performance. In the evaluations through 1998 separate tests were included with segments of three different durations: 30 seconds, 10 seconds, and 3

seconds. The results were consistently the expected ones, with improved performance with increasing duration.

In the 1999 evaluation, the one-speaker segments were of variable duration, while the two-speaker segments were all of about one minute in duration. For one-speaker detection, it was found that for segments of at least 15 seconds in duration, performance did not vary with duration. Thus a duration of 10 to 15 seconds appears to be sufficient for optimal performance.

Two-speaker performance in 1999 was analyzed as a function of true speaker duration within target trial test segments. Here performance was found to improve with increasing true speaker duration up to about 35 seconds, and then to not vary appreciably with increasing duration. Thus there was benefit in having speech from the true speaker make up a greater part of the one-minute total duration up to a point where the true speaker speech constituted a clear majority of the total speech.

### 8. Training/Test Handset Match

It has long been recognized that a major factor affecting speaker recognition performance with telephone speech is handset variation. Performance is enhanced if the handset differences between training and test do not need to be taken into account. Many potential applications, however, require effective performance where a speaker may use different handsets on different occasions..

The Switchboard Corpora rely on speakers to use their home handsets or other handsets of their choosing, and do not directly document the handsets used. They do, however, record the phone numbers for both sides of each conversation. The NIST evaluations, therefore, have relied on this phone number information to infer whether

Name	Speaker data included	
One-session	Two minutes from one conversation	1996-98, 2000
Two-session	One minute from each of two conversations from same phone number	1996-99
Two-session-all	All data from the two-session conversations	1998
Two-handset	One minute from each of two conversations from different phone numbers	1996-97

or not the training and test handsets for each target trial are (probably) identical. Each previous evaluation has included numerous target trials of each of these types. In different years the primary emphasis has been on same number or on different number target trials.

Note, however, that by the nature of the corpora, different speakers essentially never use a common handset. Thus an evaluation emphasizing same number target trials is inherently unreasonable easy, and could be viewed as constituting channel recognition as much as speaker recognition. The 2000 evaluation will have all its target trials be different number trials.

# 9. Telephone Handset Microphone Types

Conventional telephone handsets, at least in the U.S., fall into two main classes based on the type of microphone used. Most home phone handsets have electret type microphones, while many public phones use the older carbon-button type microphone. The differences in handset type affect performance in two ways. Performance is generally superior when electret handsets are used. Performance is, however, also degraded by a mismatch of handset type between a speaker's training and test conversations.

MIT-Lincoln Laboratory in Lexington, MA, a participant in the NIST evaluations, has made available to NIST its automatic handset type detector. This allows automatic determination, with high though less than perfect accuracy, of the handset microphone types of all conversation sides. In recent evaluations, this information has been made available to systems to use, if they wish, in recognition. It has also enabled subsequent analysis of performance results as a function of training and test handset microphone types.

Since electret type handsets are the more common, the more recent NIST evaluations have placed primary emphasis on trials involving this microphone type.

# 10. Conclusion

The NIST evaluations have become noted worldwide for contributing to state-of-the-art research in the core problems of speaker recognition. In 1999 12 sites from 8 countries offered results from 30 different systems in the evaluation. Conducting these evaluations involves a number of difficult and sometimes contentious issues. NIST intends to continue its annual evaluations in the year 2000 and hopefully beyond. Suggestions on improving the evaluations, particularly pointers to additional appropriate sources of data, are welcome. And new participants continue to be welcome.

# 11. References

- Doddington, G., 1998, Speaker Recognition Evaluation Methodology An Overview and Perspective -, *RLA2C*, Avignon, 60-66.
- Doddington, G., et al., 2000, The NIST Speaker Recognition Evaluation Overview, Methodology, Systems, Results, Perspective -, to appear in Speech Communication Journal.
- LDC, 2000, University of Pennsylvania, 3615 Market Street, Suite 200, Philadelphia, PA 19104-2608. web site: <a href="http://morph.ldc.upenn.edu">http://morph.ldc.upenn.edu</a>.

- Martin, A., et al., 1997, The DET curve assessment of detection task performance, *Proceedings EuroSpeech* 4:1895-1898.
- Martin, A. and Przybocki, M., 2000 The NIST 1999 Speaker Recognition Evaluation - An Overview, to appear in DSP Journal.
- NIST, 2000, Spoken Natural Language Processing Group speaker recognition web page: <a href="http://www.nist.gov/speech/spkrinfo.htm">http://www.nist.gov/speech/spkrinfo.htm</a>.
- Ortega-Garcia, J., et al.,2000, 'AHUMADA': A Large Speech Corpus in Spanish for Speaker Characterization and Identification, to appear in *Speech Communication*.
- Przybocki, M. and Martin, A., 1999, The 1999 NIST Speaker Recognition Evaluation, Using Summed Two-Channel Telephone Data for Speaker Detection and Speaker Tracking, *Proceedings Eurospeech*, 5:2215-2218.