Odyssey Text Independent Evaluation Data

Mark A. Przybocki, Alvin F. Martin

National Institute of Standards and Technology
Gaithersburg, MD 20899 USA
mark.przybocki@nist.gov alvin.martin@nist.gov

Abstract

We discuss the text-independent data supplied for the *2001: A Speaker Odyssey* evaluation track. We cover the data creation and selection process, and we present results restricted to the Odyssey test set for participating systems in the 2000 NIST Speaker Recognition Evaluation.

1. Introduction

NIST (The National Institute of Standards and Technology) has coordinated evaluations of text-independent speaker recognition using conversational telephone speech over the past six years [1]. The primary source of conversational telephone speech data has been the Switchboard Corpus [3], and more recently the various phases of the Switchboard II Corpus.

With suitable sources of data in limited supply, NIST has found the Switchboard databases to be an invaluable resource for speaker recognition research. Conversational telephone speech is a realistic form of data to test the state-of-the-art in text-independent speaker recognition. The challenges presented by this data include limited bandwidth, channel noise from various sources, the use of different microphones, recordings from different locations, and recordings collected over a period of time.

For the Odyssey text-independent evaluation track NIST selected a subset of the data that was used in the 2000 NIST Speaker Recognition evaluation [2]. Below we define and document the process of generating and selecting this evaluation data and the subset thereof that was used for the Odyssey evaluation track. We then present some results for actual competitive systems, limiting the results to this subset.

2. Evaluation Data Sources

In order to maximize the use of the available Switchboard databases [3], the 2000 NIST Speaker Recognition evaluation recycled and re-segmented speech data from Switchboard II phases 1 and 2, data which was used in the 1997 and 1998 evaluations. Switchboard II phase 3 data, as used in the 1999 evaluation, was made available to participants as development data.

Switchboard II phase 1 is a collection of about 3600 recorded telephone conversations from participants mainly in the Northeastern United States. They are conversations between two adults, usually college students, who were given a suggested topic, but who were also given permission to deviate from the topic. They converse for 5 minutes. There are about 660 speakers, each participating in an average of 11 calls. To aid in speaker recognition research, speakers were required to initiate each of their calls from a different telephone, and each speaker was only allowed to receive and initiate one call per day.

Switchboard II phase 2 had the same collection protocols as phase 1. This collection contains about 4600 conversations from about 680 speakers, each averaging just over 13 calls. The collection of phase 2 was concentrated in the Midwestern United States.

3. Training Data Generation

The target speakers in the NIST 2000 evaluation consisted of all speakers in the available corpora (Switchboard-2 phases 1 and 2) who initiated at least one call in which he or she spoke for at least two minutes. The training data then consisted of two minutes of speech from a single side of a conversation initiated by the speaker. The collection protocol thus implied that all other conversation sides of the speaker, from which test segment data might be drawn, would use a handset different from that used for the training data. This maximized the amount of available different handset test data, which the evaluation sought to emphasize.

The creation of the training segments from selected conversation sides consisted of the following steps:

1. the entire conversation was processed with publicly available echo canceling software [4]
2. the channel corresponding to the speaker of interest was separated using NIST's `w_edit` program, part of NIST's SCLITE speech recognition scoring package [5]
3. time intervals that correspond to the speaker's speech were automatically detected with the use of an energy detector [6]
4. time intervals were selected starting from the tail end of the conversation and totaling 115 to 125 seconds in duration and spliced together with NIST's `w_decode` program

This process thus removed areas of silence and yielded an approximately equal duration of training speech for each speaker.

4. Test Data Generation

Conversations for which either side had been used to generate training data were not used as sources of test data. Each side of all remaining conversations was used to generate a test segment. A random one-minute interval of each conversation side was identified. The speech corresponding to each speaker in this minute was spliced together to form the two test segments. Thus the test segment durations varied from close to zero seconds to almost a minute, but the great majority were between 15 and 45 seconds.

Note that there could be, and in fact were, a limited number of instances of target speakers who did not speak in any test segments, and of test segments whose speaker was not a target with defined training data.

5. Test Trial Definitions

An index file defines the set of evaluation trials. Each trial consists of a designated target speaker with training data and a designated test segment. A trial where the target speaks in the test segment is a true-speaker or target trial; one where someone else is the test-segment speaker is an impostor or non-target trial.

The system must decide for each trial if the given target speaker is speaking in the test segment, by providing both an actual decision (true/false) and a likelihood score. All trials must be performed independently of each other, and the likelihood scores must all use a common scale, with larger values indicating greater likelihood that a trial is in fact a true-speaker trial. This permits the generation of a full range of operating points for the system being tested.

In general for the main task of the 2000 evaluation, each test segment was used in eleven trials. The actual speaker was the target speaker in one of these. Thus there was about a ten-to-one ratio of impostor to true speaker trials. All impostor trials involved two speakers of the same sex.

6. Odyssey Test Set

While the 2000 NIST Speaker Recognition evaluation data set was distributed on 8 CD-ROMS and included 1003 speakers and 66,572 trials, there was a desire to create an evaluation kit for the Odyssey text-independent evaluation track that could be distributed on single CD-ROM.

For each evaluation task NIST defines a condition of primary interest, and sites often tune their systems to optimize performance on trials satisfying this condition. For the 2000 evaluation main task the primary condition was defined to be those trials where the test segment duration was in the 15-45 second range, and where both the test segment and the target speaker training data came from conversation sides that utilized an electret type microphone in the telephone handset. The importance of microphone type for performance results has been established in previous evaluations, and NIST utilized software from Lincoln Laboratory [7] to automatically determine this type for all conversation sides used in the evaluation.

For the Odyssey text-independent track it was decided to use all the male trials of the evaluation satisfying the primary condition. This included 417 speakers, 1,933 test segments, and 20,728 trials. By compressing this data using `shorten` [8], this entire set of data could fit on a single CD-ROM.

7. 2000 System Performance

The official performance measure for the NIST evaluations has been a weighted average (denoted C_{DET}) of the miss and false alarm error rates as defined in figure 1.

Detection Cost Function

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

Parameter Values

Cost of a miss:	$C_{Miss} = 10$
Cost of a false alarm:	$C_{FalseAlarm} = 1$
Probability of a target:	$P_{Target} = 0.01$
Probability of a non-target:	$P_{NonTarget} = 1 - P_{Target} = 0.99$

Figure 1: C_{DET} function with current parameters.

There are two detection costs that NIST regularly reports. The first is the C_{DET} value based on the actual decisions, which has been used as the official measure to determine the best performing system. The second detection cost is the minimum C_{DET} value over all operating points defined by the likelihood scores.

NIST rescored the systems that participated in the 2000 NIST evaluation on the Odyssey trials, i.e., on the primary condition male trials. Stacked bar charts are used to display the contributions of the two error types in a single plot. Figure 2 shows the actual decision detection costs for ten systems, plotted in order of improving performance, while Figure 3 shows the minimum detection costs for the same ten systems plotted in the same order. Note the systems ranked 7th, 8th, and

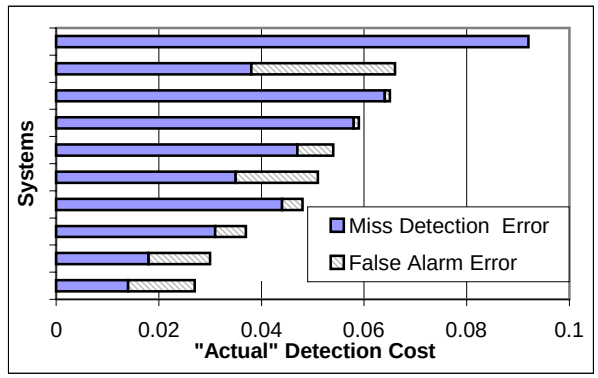

Figure 2: Actual Decision detection cost. The gray area represents the portion of error due to missed detections; the hatched area represents the portion of error due to false alarms.

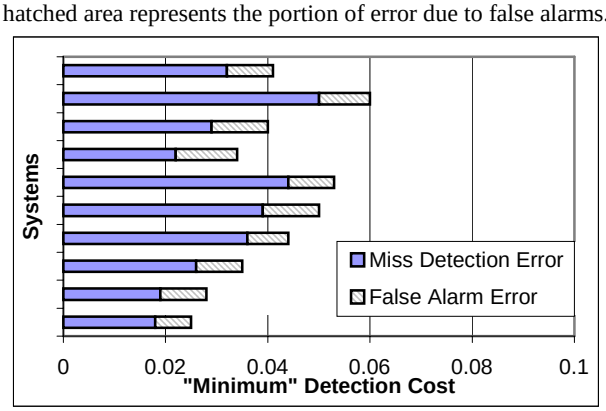

Figure 3: Minimum Detection Cost over all operating points for the same ten systems as in Figure 2. An optimal choice of thresholding value for actual decisions would result in Figure 2's values approaching what is shown here.

10th based on actual C_{DET} cost ranked in the top six on minimum C_{DET} cost. This suggests that these three systems had relatively less well chosen likelihood threshold settings (for the trials under consideration) for determining the actual decisions than other systems.

Figure 4 presents a DET plot for these same ten systems. A DET curve [9] displays all operating points for a given system, with normal deviate scales on both axes. The curves also identify the two special operating points with special characters: a circle representing the minimum C_{DET} point and a diamond representing the actual decision C_{DET} point.

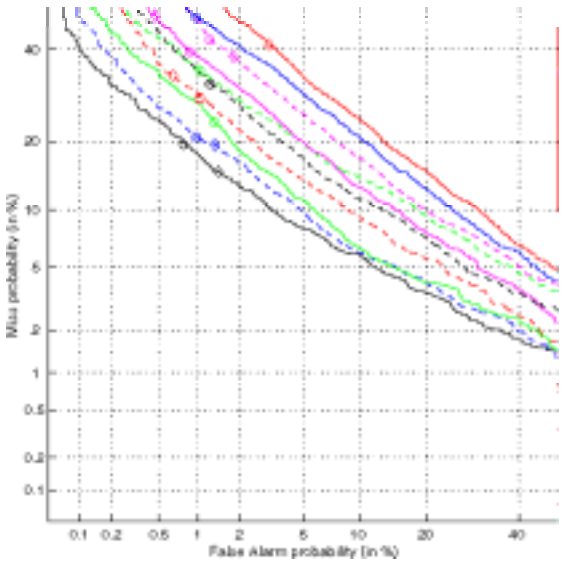

Figure 4: DET plot of ten NIST 2000 Speaker Recognition participating systems processing the male primary condition data

Figure 5 shows for one system the difference in performance between processing all the primary condition trials, and the subset selected for use in the Odyssey evaluation track. This system was typical of most in this evaluation in having slightly better performance on the male trials than on the female trials.

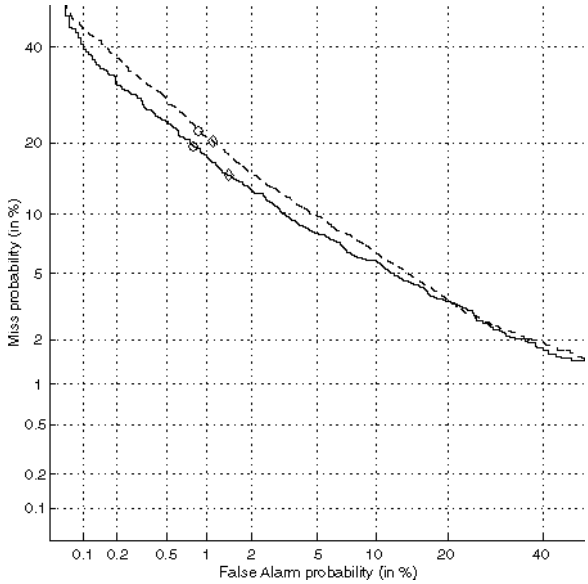

Figure 5: One system, processing all primary condition data (dashed) and the Odyssey subset (solid).

References

[1] A. Martin and M. Przybocki, "The NIST Speaker Recognition Evaluations: 1996-2001", Proc. Odyssey Workshop, June 2001

[2] NIST 2000 Speaker Recognition Evaluation, http://www.nist.gov/speech/tests/spk/2000/index.htm

[3] Switchboard Databases are available from the LDC. http://www.ldc.upenn.edu/

[4] The echo canceling software was provided by ISIP, the LDC provided a Perl wrapper script, for more information see: http://www.nist.gov/speech/tests/ctr/h5e_97/echocan.htm

[5] NIST's SCLITE speech recognition scoring is available from: http://www.nist.gov/speech/tools/index.htm

[6] The energy detector is embedded in NIST SPeech Quality Assurance (SPQA) package, available from: http://www.nist.gov/speech/tools/index.htm

[7] D. Reynolds, "HTIMIT and LLHDB : Speech Corpora for the Study of Handset Transducer Effects", Proc. ICASSP97

[8] SHORTEN: Simple lossless and near-lossless waveform compression, Tony Robinson, Technical report CUED/F-INFENG/TR.156.

ftp://svr-ftp.eng.cam.ac.uk/pub/reports/robinson_tr156.ps.Z

[9] A. Martin et al., "The DET Curve in Assessment of Detection Task Performance", Proc. EuroSpeech 1997, Vol 4, pp. 1895-1898.