THE 1999 NIST SPEAKER RECOGNITION EVALUATION, USING SUMMED TWO-CHANNEL TELEPHONE DATA FOR SPEAKER DETECTION AND SPEAKER TRACKING

Mark A. Przybocki *Alvin F. Martin*

National Institute of Standards and Technology, 100 Bureau Dr. Stop 8940, Gaithersburg, MD 20899, USA
mark.przybocki@nist.gov http://www.nist.gov/speech/spkrinfo.htm

ABSTRACT

The 1999 NIST Speaker Recognition Evaluation encompassed three tasks: one-speaker detection, two-speaker detection, and speaker tracking. All tasks were performed in the context of conversational telephone speech. The one-speaker task used single channel mu-law data; the other tasks used summed two-channel data. Twelve sites from the United States, Europe, and India participated in the evaluation. Performance was measured by a decision cost function and compared among systems and test conditions via DET Curves. Performance factors examined include segment duration, degradation resulting from the presence of a second speaker, sex mix of the two-speaker segments, matched or mismatched between training and test handsets, and the variation in handset type.

1. INTRODUCTION

NIST has coordinated speaker recognition evaluations using conversational telephone data for the past five years [6]. These evaluations have played an important role in directing the research efforts and calibrating the technical capabilities of the core technology. These evaluations are designed to be fully supported and accessible to all researchers who are developing text-independent speaker recognition systems.

In 1998 and in previous evaluations, systems processed test segments that contained speech from only one speaker, and the task was to determine whether or not the test segment contained speech of a hypothesized speaker. The 1999 evaluation, while continuing the single speaker evaluation of previous years, also focused on the presence of more than one speaker in the test stream.

During the summer of 1998, a development evaluation was conducted by NIST. In this evaluation, systems processed test segments that had the added complexity of more than one speaker present. The detection task required, given a test segment, to identify whether or not a hypothesized speaker is speaking anywhere in the test segment. The tracking task required, given a test segment, to identify the exact intervals (if any) where a hypothesized speaker was speaking.

The summer development evaluation was a success in that it identified challenging tasks that fit the framework of the NIST speaker recognition evaluations. These tasks were included in the official 1999 NIST Speaker Recognition Evaluation as the two-speaker detection task and the speaker tracking task. For some applications, such as in information extraction and retrieval from a recorded meeting, these may be viewed as the more realistic speaker verification tasks.

A detailed description of how each task was implemented is described in the official 1999 Speaker Recognition Evaluation Plan [1]. We discuss here implementation of the evaluation in brief, to aid understanding of these preliminary results.

2. THE EVALUATION

The 1999 NIST speaker recognition evaluation had three distinct tasks, one-speaker detection, two-speaker detection and speaker tracking. Each task was evaluated separately.

For all three tasks the formal evaluation measure was the detection cost function, defined as a weighted-sum of the miss and false alarm error probabilities:

$$C_{Det} = C_{Miss} \times P_{Miss|Target} \times P_{Target} + C_{FalseAlarm} \times P_{FalseAlarm|NonTarget} \times P_{NonTarget}$$

The parameters of this cost function are the relative costs of detection errors, C_{Miss} and $C_{FalseAlarm}$, and the *a priori* probability of the target, P_{Target}. The following parameter values were used for both the detection and the tracking tasks:

$$C_{Miss} = 10;\ C_{FalseAlarm} = 1;\ P_{Target} = 0.01;$$
$$P_{NonTarget} = 1 - P_{Target} = 0.99$$

2.1 Speaker Detection

For the one-speaker detection task, each test segment was to be judged as true or false for each of eleven hypothesized speakers, one of which was the true speaker of the test segment. In addition to this binary detection decision a decision score was also required. These decision scores are used to produce detection error tradeoff (DET) curves, in order to see how missed detections may be traded off against false alarms. DET curves provide a convenient way to view the performance of several systems (or several conditions) in one graph [2].

For the two-speaker detection task, each test segment was to be judged as true or false for each of twenty-two hypothesized speakers, two who were truly present in the test segment, twenty who were not. As in the one-speaker detection task, a decision a decision score was also required.

2.2 Speaker Tracking

For the speaker tracking task, the system had to determine the intervals where each of four hypothesized speakers *(4 of the 22 used in the detection task, 2 being actual speakers, 2 being imposters)* are speaking in the test segment. As in the detection task, for each system dependent interval a hard decision and a decision score is given. System output decisions were compared with a reference answer key. NIST used an automatic speech detector to establish the reference intervals of speech for each channel of the test segments. With this comparison, NIST measured the miss and false alarm rates for the speaker tracking task according to the following computation:

$$P_{Miss} = \int_{\text{Target Speech}} \delta\,(D_t,F)dt \;\Big/ \int_{\text{Target Speech}} dt \qquad P_{FalseAlarm} = \int_{\text{Impostor Speech}} \delta\,(D_t,T)dt \;\Big/ \int_{\text{Impostor Speech}} dt$$

Where: D_t = the system output as a function of time
$\delta(x,y) = \{\ 1 \text{ if } x{=}y,\ \ 0 \text{ otherwise }\}$

3. EVALUATION DATA

The data for this evaluation was drawn from the Switchboard-2 Phase-3 corpus, available from the Linguistic Data Consortium [3]. This corpus consists of about 2,700 recordings of five-minute telephone conversations between just over 600 English-speaking subjects. The majority of the subjects were recruited from the South of the United States. This was done in an attempt to obtain speakers with similar dialects. Each speaker was encouraged to participate in five calls from a single phone line, and in five calls from five different phone lines.

3.1 Evaluation Training Data

Training data was provided for each speaker in the corpus who participated in two separate conversations from the same line number. There was a total of two-minutes of training data for each model. The training data originated from two separate conversations (one minute from each). Successive speech segments were concatenated together, removing silences. The nominal duration of 60 seconds varied slightly to allow whole segments to be included. NIST manually verified that each training segment was a fair representation of the target speaker.

Training segments were provided for 230 male speakers, and for 309 female speakers. 972 conversations were used for training, and therefore were not available for evaluation data.

3.2 Evaluation Test Data

All remaining conversations were used for constructing the one-speaker and two-speaker evaluation test segments. Only one two-speaker test segment was taken from a single conversation. These test segments, one minute in length, were extracted from the end of a conversation. Both channels were summed to produce the test segment. Unlike the training data, areas of silence were not removed. These two-speaker test segments were used for both tasks, detection and tracking. However, the tracking task was limited to 1000 of these segments.

Two corresponding one-speaker test segments were created from the separate channels of each two-speaker segment, with the speech intervals of the particular speaker concatenated together.

4. EVALUATION RULES

Every evaluation participant was required to undertake the full test for at least one of the three evaluation tasks: 1-speaker detection, 2-speaker detection, or speaker tracking.

Each decision was to be based only upon the specified test segment and the hypothesized speaker. Normalization over multiple test segments or multiple target speakers, was not allowed. Participants were not allowed to use the evaluation data for impostor modeling. The use of transcripts for training the target speakers was not allowed.

Although the sex of the hypothesized speaker was known, knowledge of the sex mixture of the test segments was not allowed, unless determined by an automatic process.

Knowledge of the handset type (electret or carbon-button, determined by using an automatic handset detector from MIT Lincoln Labs [4]) was allowed for the 1-speaker task but not for the 2-speaker task unless determined by an automatic process.

Experimental interaction with the evaluation data was not allowed. This included listening to either the training or test segments.

The corpus from which the evaluation was drawn, namely the Switchboard-2 Phase 3 corpus, could not be used for any system training or R&D activities relating to this evaluation.

5. PARTICIPANTS

Participation in the NIST Speaker Evaluations has consistently grown through the years. In the past three years over 18 different research sites have participated in our evaluations. In this year's evaluation the list of participants included: *Dragon Systems, Ecole Nationale Superieure des Telecommunciations, Ensigma Ltd, EPFL-RMA-VUT-ENST collaboration, FPMS-RICE collaboration, IIT Madras, Institut Dalle Molle d'Intelligence Artificielle Perceptive, Institut de Recherche et Informatique et Systemes Aleatoires, MIT Lincoln Labs, Nijmegen University, Oregon Graduate Institute, Universite d'Avignon.*

Eleven participants submitted on-time results for the 1-speaker detection task, four participants submitted on-time results for the 2-speaker detection task and five submitted on-time results for the speaker tracking task. All results were accompanied by a complete system description.

6. OVERALL RESULTS

6.1 One-Speaker Detection Results

The 1-speaker detection task consisted of 37,620 decisions. NIST identified the primary condition of interest as target trials that were:

- Different line tests.
- Both training and test segments from electret microphones.
- Test Segment durations of 15-45 seconds.

Impostor trials were also restricted to the same conditions (all impostor trials are different line tests).

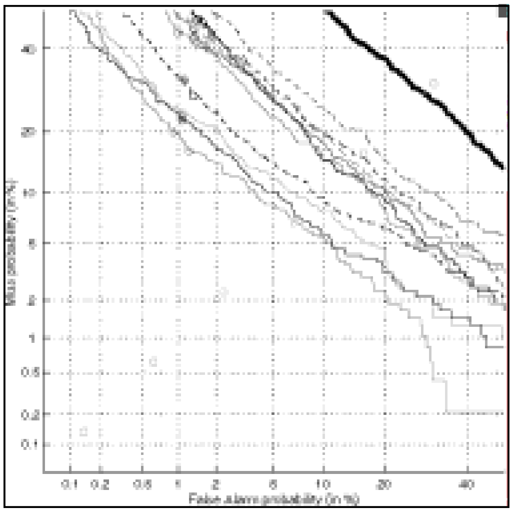

Figure 1: Results from the 1-speaker detection task. The 11 primary systems are shown. For each system the actual decision point is plotted with a circle, the minimum C_{Det} point is plotted with a diamond.

As Figure 1 shows, the performance varies widely across sites. There appears to be a front running pack of the four top-performing systems, than a middle cluster followed by one system which shows particularly poor results.

6.2 Two-Speaker Detection Results

The primary condition for the two-speaker detection task was:

- Different line tests
- All models and both sides of test segments from electret microphones
- Test segment side duration of 15-45 seconds

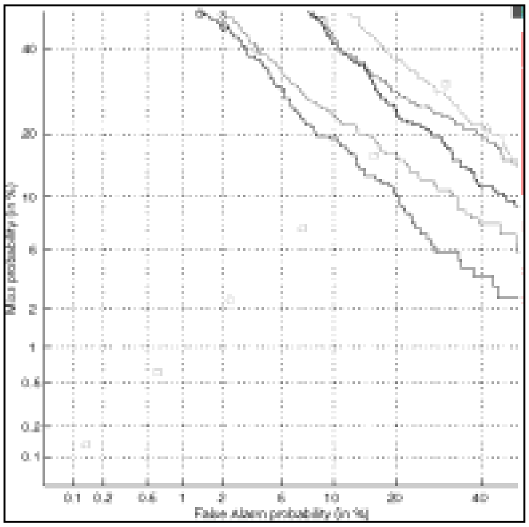

Figure 2: Results from the 2-speaker detection task. The 5 primary systems are shown. For each system the actual decision point is plotted with a circle, the minimum C_{Det} point is plotted with a diamond.

Figure 2 presents two-speaker detection primary results for systems from five sites. Again there is considerable variation in performance among systems.

6.3 Speaker Tracking Results

Primary results for the speaker tracking task included all of the tracking segments and hypothesized speakers included in the test. Performance results for systems from 5 sites are shown in Figure 3. Note that the results are relatively and fairly uniformly, poor. This det plot has the (50%, 50%) point of random performance in the center of the chart. This is clearly a difficult task, and further consideration is needed of how best to evaluate it.

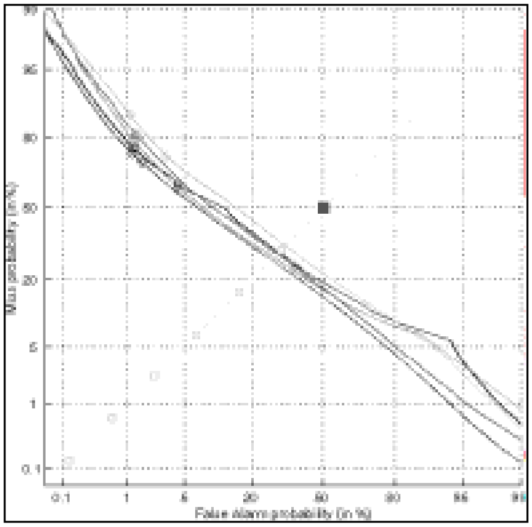

Figure 3: Results from the speaker tracking task. The 5 primary systems are shown. For each system the actual decision point is plotted with a circle, the minimum C_{Det} point is plotted with a diamond.

7. COMPARATIVE RESULTS

7.1 One-Speaker vs. Two-Speaker Detection Results

Two-speaker detection is clearly a harder task than one-speaker detection. Figure 4 indicates just how much harder this task is. Results are shown for three different systems with the same tests included in each task, namely all the tests corresponding to the two-speaker primary conditions which were also one-speaker tests.

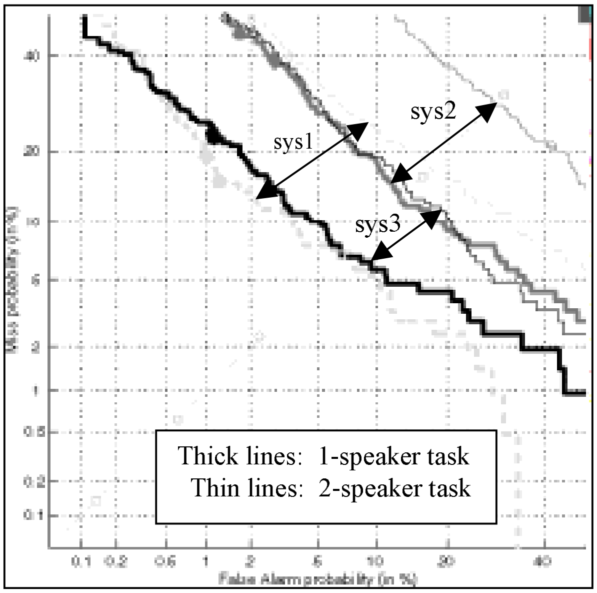

Figure 4: One-speaker and two-speaker detection performance compared for three systems. The two-speaker primary tests are included in all cases.

7.2 Handset Match and Type

Target trials may be either same number or different number according to the phone line (and presumably the telephone handset) used in training the model and used in the test segment. Note that this is a condition on target trials only; for impostor trials the two phone lines are always different.

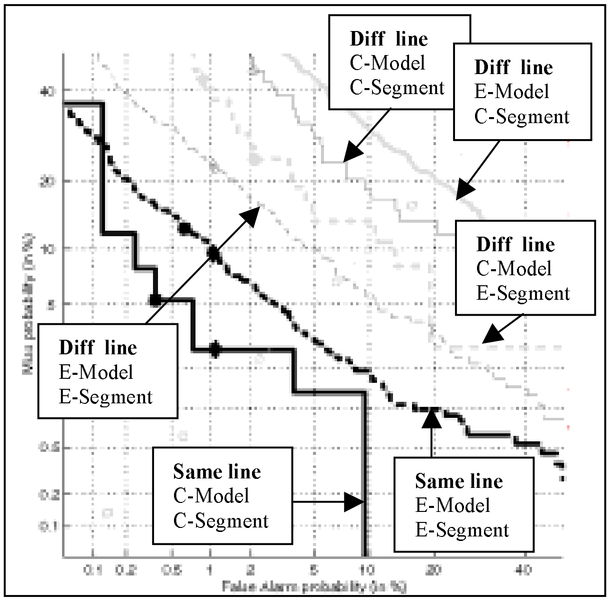

Figure 5: One-speaker detection performance for same and different number target trials and for electret and carbon-button handset types of model and segment. Results for one system are shown.

For different number trials we may distinguish four handset type conditions:

(1) model is electret, segment is electret
(2) model is carbon, segment is electret
(3) model is electret, segment is carbon
(4) model is carbon, segment is carbon

Figure 5 shows performance for one system for the six resulting conditions (2 from same number, 4 from different number). For this system and for most systems the following may be noted:

- Performance is far superior for same number tests compared to different number tests
- For different number tests, performance is best when both the model training and the test segment are electret
- For different number tests, performance is better when the test segments are electret than when they are carbon-button.

7.2 Test Segment Duration

In previous NIST evaluations the one-speaker segments were all of approximately 3, 10, or 30 seconds in duration, and performance was, as expected, improved as segment duration increased [5]. This year's one-speaker tests varied in duration from 0 to 60 seconds since the segments corresponded to the one-speaker portion of the one-minute two-speaker segments. Figure 6 examines performance as a function of test segment duration. For a typical system it may be noted that the shortest segments (under 15 seconds) had inferior performance, but duration mattered little for segments of more than 15 seconds.

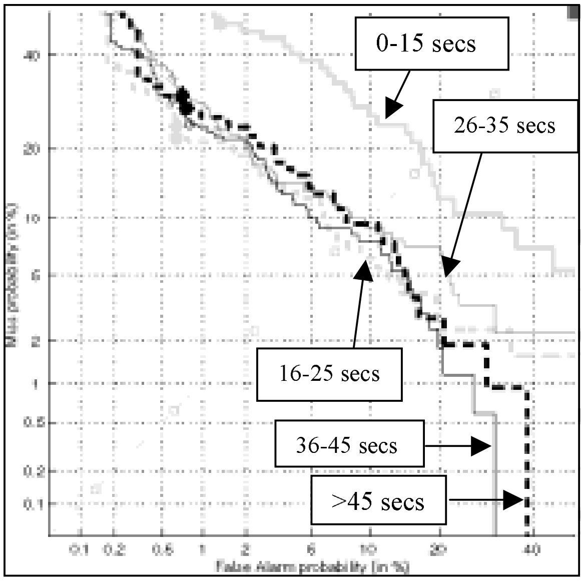

Figure 6: Performance comparison of one-speaker detection by duration of test segments (target and impostor) for one system.

7.3 Sex Mixture

There were no cross-sex tests (male model with female speech segment or vice versa) in this evaluation, and it is of interest to compare performance on males and females. For the two-speaker tests there is the added factor of whether only one or both of the segment speakers are of the same sex as the model. Figure 7 shows results with respect to these conditions for two fairly typical systems. One system performed better on male (model) tests, while the other did better on female tests. There was no overall sex preference in performance. But performance was generally better for the mixed sex tests where one of the test segments differed in sex from that of the model.

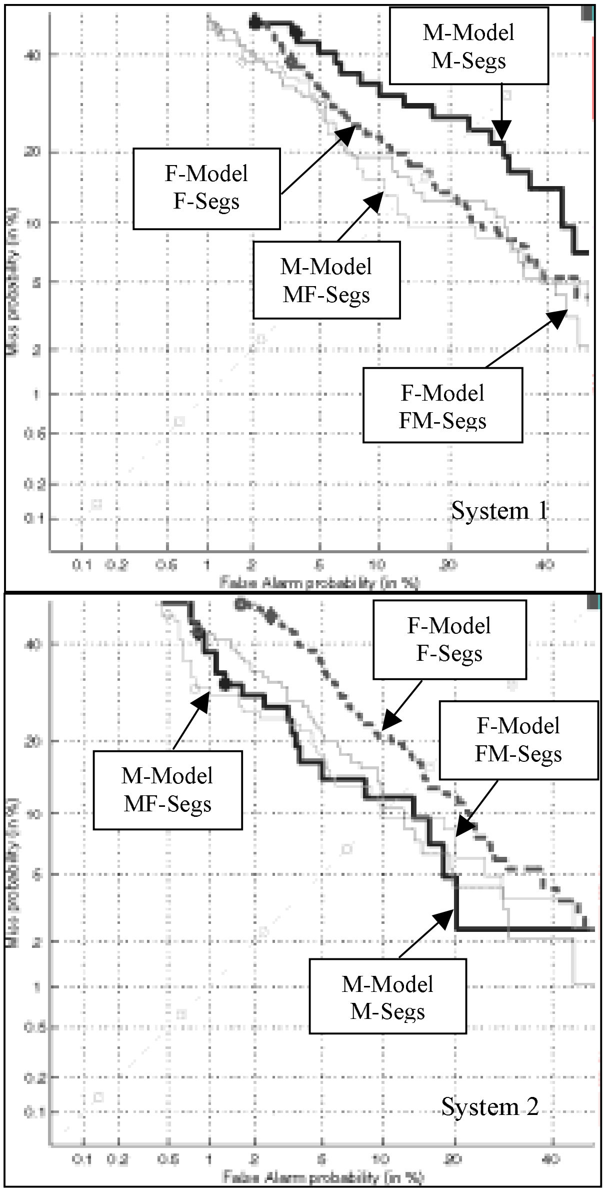

Figure 7: Two speaker detection performance according to the gender mix of the model and test segments (target and impostor) for two different systems.

8. REFERENCES

[1] Martin A., et al. "The 1999 NIST Speaker Recognition Evaluation Plan", NIST web-site: www.nist.gov/speech/spk99/spk99plan.html, February 1999.

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

[3] Linguistic Data Consortium, http://www.ldc.upenn.edu.

[4] Reynolds, D. "HTIMIT and LLHDB: Speech Corpora for the Study of Handset Transducer Effects,," ICASSP pp. 1535-1538 April 1997.

[5] Przybocki M., Martin A., "NIST Speaker Recognition Evaluation - 1997", RLA2C proceedings pp. 120-123, 1998.

[6] Doddington G., et al. "NIST Speaker Recognition Evaluation - Overview, Methodology, System, Results, Perspective -", to appear in Speech Communication.