NIST 2008 Speaker Recognition Evaluation: Performance Across Telephone and Room Microphone Channels

Alvin F. Martin, Craig S. Greenberg

National Institute of Standards and Technology, Gaithersburg, Maryland, USA

alvin.martin@nist.gov, craig.greenberg@nist.gov

Abstract

We describe the 2008 NIST Speaker Recognition Evaluation, including the speech data used, the test conditions included, the participants, and some of the performance results obtained. This evaluation was distinguished by including as part of the required test condition interview type speech as well as conversational telephone speech, and speech recorded over microphone channels as well as speech recorded over telephone lines. Notable was the relative consistency of best system performance obtained over the different speech types, including those involving different types in training and test. Some comparison with performance in prior evaluations is also discussed.

Index Terms: speaker recognition evaluation, speaker detection, NIST evaluations

1. Introduction

The 2008 NIST Speaker Recognition Evaluation (SRE08) was conducted during the spring of 2008. Like its preceding evaluations since 1996, it concentrated on the task of speaker detection. Each evaluation test condition consisted of a (long) sequence of individual trials, where each trial consisted of a target speaker, defined by one or more segments of training speech by the target, and a test segment of speech by an unknown speaker. For each trial an evaluation system must provide a decision (‘T’ or ‘F’) on whether or not the unknown is the target, and a likelihood score indicating degree of belief that “T” is the correct decision.

Prior evaluations concentrated on conversational telephone speech, at least in the core condition required of all participants. The 2008 evaluation, in contrast, required all systems to process trials involving conversational speech recorded over multiple room microphones as well as such speech recorded over telephone lines.

Further information about SRE08 is available on its web page (www.itl.nist.gov/iad/mig/tests/sre/2008/index.html). The page has links to the official evaluation plan provided to participants and to a page summarizing the evaluation performance results, which in turn links to a brief slide presentation on the evaluation and its results.

Section 2 discusses the evaluation data and core test condition, while Section 3 describes the participants. Section 4 presents some performance results for one evaluation system, and Section 5 offers a comparison of SRE06 and SRE08 results one site. Sections 6 describes briefly the follow-up evaluation to SRE08; Section 7 notes plans for SRE10; and Section 8 is a brief summary.

Descriptions of the NIST SRE’s for years prior to 2008 may be found in [1, 2, 3, 4, 5, and 6]. Other aspects of SRE08 are discussed in [7].

2. Data Source and Core Test Condition

SRE08 utilized speech from the Mixer 3, Mixer 4, and Mixer 5 Corpora collected by the Linguistic Data Consortium [8, 9]. These corpora consist, respectively, of recorded telephone conversations of paired speakers who do not know each other, recordings of some of these telephone conversations over multiple room microphones, and recordings of interview sessions, involving both conversational and read speech, over multiple room microphones. Mixer 3 includes many conversations involving bilingual speakers some of whose conversations were in a language other than English.

The core test condition involved selected segments from each of these corpora. The training segments were of two types:

- One two-channel telephone conversational excerpt of approximately five minutes total duration, with the target speaker channel designated (Mixer 3)
- A room-microphone recorded conversational segment of approximately three minutes total duration involving the target speaker and an interviewer (Mixer 5)

These durations were chosen in the expectation that both would average about two and a half minutes of speech by the target speaker. The test segments were of three types:

- One two-channel telephone conversational excerpt of approximately five minutes total duration, with the target speaker channel designated (Mixer 3)
- A similar such telephone conversational excerpt recorded over a room microphone (Mixer 4)
- A room-microphone recorded conversational segment of approximately three minutes total duration involving the target speaker and an interviewer (Mixer 5)

The numbers of target and non-target trials by data source for the core test, along with the numbers of models and test segments by source are specified in Table 1.

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Mixer 3</th>
<th>Mixer 4</th>
<th>Mixer 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>2573 segs</td>
<td>1460 segs</td>
<td>2344 segs</td>
</tr>
<tr>
<td>Mixer 3</td>
<td>3832/33218</td>
<td>1472/6982</td>
<td>2500/4850</td>
</tr>
<tr>
<td>Mixer 5</td>
<td>1105/10636</td>
<td>Not tested</td>
<td>11540/22641</td>
</tr>
</tbody>
</table>

In past evaluations, a “common condition” was defined as a subset of the core test that would be used primarily for system comparisons. In SRE08, because of the multiple speech data types used in the core test, eight common conditions for system comparison were defined as follows:

1. All trials with Mixer 5 (interview) speech in training and test
2. All trials with Mixer 5 (interview) speech from a common microphone in training and test
3. All trials with Mixer 5 (interview) speech from distinct microphones in training and test
4. All trials with Mixer 5 (interview) training speech and Mixer 3 (telephone) test speech
5. All trials with Mixer 3 (telephone) training speech and Mixer 4 test speech
6. All trials with Mixer 3 (telephone) speech in training and test
7. All trials with English language Mixer 3 (telephone) speech in training and test
8. All trials with English language Mixer 3 (telephone) speech spoken by a native U.S. English speaker in training and test

3. Participants

There has been a steady increase in the number of participating research organizations over the course of the NIST SRE’s, and SRE08 was no exception. Sites were encouraged to participate either by themselves and/or in teams with other participants, allowing algorithms and systems to be shared and combined, and permitting some sites to concentrate on particular aspects of system design. There were numerous such combinations in 2008, with 46 different participating organizations and also, coincidently, 46 different sites or combinations of sites submitting one or more evaluation systems. There were 107 systems in all (each of the 46 was asked to designate a primary system) and a total of 246 different test condition/system combinations. (Note that all systems included the core test condition.)

The participants in SRE08 included companies, universities, and government research organizations from around the world. The full list may be found on the SRE08 web site referenced above. Table 2 lists the numbers of participating sites by country.

<table>
<thead>
<tr>
<th>Country</th>
<th>Number of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>2</td>
</tr>
<tr>
<td>Canada</td>
<td>1</td>
</tr>
<tr>
<td>China</td>
<td>9</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>2</td>
</tr>
<tr>
<td>Finland</td>
<td>1</td>
</tr>
<tr>
<td>France</td>
<td>5</td>
</tr>
<tr>
<td>Germany</td>
<td>1</td>
</tr>
<tr>
<td>Israel</td>
<td>1</td>
</tr>
<tr>
<td>Italy</td>
<td>2</td>
</tr>
<tr>
<td>Lebanon</td>
<td>1</td>
</tr>
<tr>
<td>Lithuania</td>
<td>1</td>
</tr>
<tr>
<td>Mexico</td>
<td>1</td>
</tr>
<tr>
<td>Netherlands</td>
<td>1</td>
</tr>
<tr>
<td>Singapore</td>
<td>2</td>
</tr>
<tr>
<td>Slovenia</td>
<td>1</td>
</tr>
<tr>
<td>South Africa</td>
<td>2</td>
</tr>
<tr>
<td>Spain</td>
<td>5</td>
</tr>
<tr>
<td>Switzerland</td>
<td>1</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>1</td>
</tr>
<tr>
<td>United States</td>
<td>6</td>
</tr>
</tbody>
</table>

The 46 participating sites are listed in the web presentation noted in section 1. Note that by tradition and general agreement within the SRE research community, NIST does not associate participating site names with their performance results in public reports. However, each participating site is free to report its own SRE08 performance results in open publications if it wishes to do so.

4. Performance Results

NIST presents system performance results via Detection Error Tradeoff (DET) curves. These are variants of Receiver Operating Characteristic (ROC) curves that plot the false alarm and miss rates on normal deviate scales on the x and the y axis, respectively. The points on the resulting curve are the possible system operating points as the decision threshold is varied over the range of the likelihood scores the system supplies for each of the trials for the test condition being plotted [10].

As described in Section 2, there were eight common conditions defined for the core test. Thus eight different plots could be presented, each showing the performance of the 46 primary systems for the given condition. These eight plots, and many others, including plots for the non-core test conditions, may be viewed on the evaluation results summary web page noted in Section 1.

Figure 1 presents DET curves of some of the common condition performance results for one leading evaluation system. It includes common conditions 2, 3, 4, 5, and 7 as described in Section 2. Condition 7 is included rather than condition 6 because, as seen in past evaluations, performance results were generally considerably better over trials involving English language speech only [5, 6]. Note that the best performance curve is, perhaps unsurprisingly, for Mixer 5 same microphone trials, and that the conditions mixing telephone and microphone channels in training and test performed less well than the unmixed conditions. But since prior evaluations had emphasized telephone trials, the comparability of the microphone and telephone results, and the limited deterioration seen in the mixed conditions was viewed as a positive and somewhat unexpected outcome. This is discussed further in the next section.

5. Comparison With SRE06

The Mixer3 and Mixer4 speech data used in SRE08 were collected using similar protocols to those for corresponding speech data used in SRE06. Nevertheless, in comparing the performance results of successive evaluations, there is always
a question of whether one test set may have been intrinsically more difficult than the other.

One well-performing site was kind enough to run its SRE06 system on the SRE08 data. Comparison of its performance on the English language telephone trials of the two evaluations suggested that the two test sets were fairly close in difficulty, with the 2008 one perhaps slightly harder.

Figures 2 and 3 present this site’s 2006 and 2008 results on trials with test segments from telephone conversations that were simultaneously recorded over room microphones. This was a non-core optional test condition in 2006. In both figures the solid black curve represents trials with telephone test (and training) segments, while the other curves represent test segments recorded over a particular room microphone, always with telephone training. There is not a specific correspondence of microphones between the two figures.

It is notable that in 2006 the all-telephone trials gave distinctly superior results, while in 2008 the telephone test channel gave good results, but not always superior results to the room microphones. This outcome suggests that reasonably good cross-channel speaker detection performance was achieved in 2008.

It may also be noted that there was considerable variation in performance among the several room microphone channels tested. Some limited analysis suggested that distance of microphone placement from the speaker’s lips was one key factor affecting overall system performance.

6. Follow-up Evaluation

The Mixer 5 data used in SRE08 involved about 150 of the 300 speakers in the recorded corpus, and 8 of the 14 microphones over which the speakers were recorded. A decision was made at the evaluation workshop to invite some of the participants to take part in a follow-up Mixer 5 only evaluation conducted in the late summer of 2008.

Seven sites chose to take part. The 150 speakers and their training speech did not change from the main evaluation, but the test segment data utilized all of the recorded room microphones. Further, the set of evaluation trials was greatly expanded to include all of the three minute Mixer 5 training models of all 150 target speakers paired with all of the same-sex Mixer 5 three minute test segments in all channels. This resulted in a total number of trials in excess of six million!

This follow-up evaluation provided a rich trove of performance results. While performance curves were generated for each participating system with respect to each microphone combination, there is much interesting analysis still to be done. The full matrix set of trials will support, for example, investigation of which speakers or speaker pairs were particularly difficult to distinguish.

7. Future Plans

The next NIST speaker evaluation will take place in 2010. SRE10 will again test speaker detection in the context of both conversational telephone speech and interview speech, and utilizing both speech recorded over telephone lines and speech recorded over multiple room microphones. New evaluation data is being collected in 2009, and specific evaluation plans will be announced by the beginning of 2010. Participation will again be welcome from all interested research sites willing to complete the evaluation core test in accord with its rules and to report on the system(s) it used at the evaluation workshop. Sites interested in evaluation participation should contact NIST at speaker_poc@nist.gov.

8. Summary

The 2008 NIST Speaker Recognition Evaluation (SRE08) was in several respects the largest and most involved of the evaluation series. It had a record number of participants, a record data set size and number of trials, and a record number of common conditions within the core test. The required core test in 2008 included trials using interview speech and conversational telephone speech simultaneously recorded over room microphones as well as speech recorded over telephone lines.

Performance results were surprisingly strong for conditions involving speech recorded over room microphones, approaching those obtained with telephone speech.

A follow-up evaluation included microphone speech over an expanded set of channels and a full matrix (within sex) set of trials.

The next evaluation in the series will take place in 2010.
9. References


