Performance Factor Analysis for the 2012 NIST Speaker Recognition Evaluation

Alvin F. Martin\(^1\), Craig S. Greenberg\(^1\), Vincent M. Stanford\(^1\), John M. Howard\(^2\), George R. Doddington\(^1\), John J. Godfrey\(^3\)

\(^1\)National Institute of Standards and Technology, Gaithersburg, Maryland, USA
\(^2\)Systems Plus, Inc.
\(^3\)Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA

\{alvin.martin\, craig.greenberg\, vincent.stanford\, john.howard\, george.doddington\}@nist.gov, godfrey.jack@gmail.com

Abstract

The 2012 NIST Speaker Recognition Evaluation, held in the autumn of 2012, was designed to examine a variety of factors affecting the performance of automatic systems for speaker recognition. Here we examine, for leading systems included in this evaluation, the observed effects on performance of five such factors: the inclusion in test segment speech of environmental noise or of added synthetic noise of one of three types and one of two intensity levels, the duration of test segment speech, the number and the channel type of target speaker training sessions, the type of the microphone channel used in test segment speech, and the sex of the target speaker. This evaluation is notable for being the first in the series to include examination of the effects on performance of synthetic added noise. The greater impact of crowd noise compared to HVAC noise, and of single speaker noise compared to crowd noise is observed. Future evaluation plans are also discussed.

Index Terms: speaker recognition, speaker detection, NIST evaluation

1. Introduction

The 2012 NIST Speaker Recognition Evaluation (SRE12) \([1, 2]\) was conducted in the autumn of 2012. It was the latest in a series of NIST coordinated evaluations of this technology stretching back to 1996 \([3]\).

The most notable new feature of SRE12 was that most of the target and test segment speakers were repeated speakers from prior NIST evaluations, with all of the speech sessions from the prior evaluations available as target speaker training data. Thus, unlike in all the prior evaluations, systems were permitted to have prior knowledge of multiple speakers, including the actual speaker in most non-target trials. Some of the resulting effects of this on performance are discussed in \([4]\); here we discuss the performance effects of various other factors examined in this evaluation.

The performance results are presented in the following sections using DET curves \([5]\), as is customary in NIST speaker evaluations. It is relevant to note that for the main test conditions presented here, systems were to assume that in non-target trials the a priori probability that the speaker was one of those targets already known to the system was to be assumed to be 0.5. Because of this, the false alarm rates used in the curves presented are means of this rate over trials with known and over trials with unknown speakers. (See \([1,2]\) for further details).

Section 2 reviews briefly the test data collected and selected for SRE12, while section 3 discusses some of the evaluation conditions offered, and section 4 discusses evaluation participation. For more details on the data used and evaluation conditions offered, see \([1, 2]\). The following sections examine the several performance factors: noise, duration, training, channel, and sex are addressed in sections 5, 6, 7, 8, and 9, respectively. We summarize briefly and discuss future evaluation plans in section 10, and provide a disclaimer in section 11.

2. Evaluation Data

The test segment data for SRE12 came from the Mixer 7 and ReMix Corpora, each collected by the Linguistic Data Consortium. All data used contained English Language speech.

Mixer 7 consists of phone calls collected over telephone channels and multiple room microphone channels, and face-to-face interviews recorded over multiple room microphone channels. This corpus was also used earlier in the NIST interim assessment of speaker recognition systems \([6]\), with limited participation, conducted as part of the IARPA (Biometrics Exploitation Science & Technology (BEST) Program. \([7]\).

ReMix is a corpus solely of phone calls recorded over a telephone channel, and was collected specifically for SRE12. All of its speakers are people who participated in one of the earlier LDC Mixer Corpora, which were used for the earlier NIST SRE’s (or for BEST in the case of Mixer 7). ReMix speakers were encouraged to make up to twelve phone calls, and to do some of them in a noisy environment, such as in a vehicle, at or busy intersection, or with television or music playing in the background.

From each of the phone calls and interviews available, test segments were selected with durations of approximately 30 s, 100 s, and 300 s. Segments with additive noise (see section 3) were also generated, but only for the segments of duration 300 s.

3. Evaluation Conditions

Three system training conditions were offered in SRE12. The core (required) condition was to use all of the provided data for each target speaker. Participants were also invited to submit alternative systems for contrast in which the training for each target was limited to only the provided telephone channel data, or to only the provided microphone channel data.
Participants also had options on the total number of trials for which they might submit results. The core (required) trial set was limited to fewer than 2 million trials, while the extended trial set had a specified upper bound of 100 million. Most participants only did the core, but extended trial set results generally produced smoother performance curves, particularly when the trials were limited to those involving specific performance factors, and particularly in the very low false alarm region on the left side of DET plots.

4. Participants

SRE12 attracted participation from 58 different research sites from 24 different countries representing every continent (except Antarctica). There were 49 primary systems performing the core test (some sites participated within teams). The total number of primary and alternate systems for all the offered tests was 212.

It is NIST policy, explicitly described in the evaluation plan [8], not to publicly identify sites participating in speaker evaluations along with their system performance results. In the following sections we present results for systems that were among the leading performers in the evaluation. The SRE12 Results Page presents charts showing the performance of all primary systems for the five common conditions.

5. Noise

An effort was made in the ReMix collection to encourage participants to record some conversations with noisy conditions of various types. But beyond this necessarily limited effort, it was decided it would be useful to study the effects of noise on performance in a controlled way by adding specific types and levels of separately generated noise after-the-fact to many of the 300 s test segments.

Table 1 shows the types and levels of noise that were generated and added to interview or telephone test segments of 300 s duration. Three types of noise were included: HVAC (heating, ventilation, and air conditioning), single speaker, and crowd noise. There were two approximate SNR levels, 15 dB and 6 dB. In all, 7 of the combinations were tested.

Table 1: Types and levels of additive noise tested in SRE12 SNR.

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>6 dB</th>
<th>15 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>HVAC</td>
<td>Interview</td>
<td>Interview</td>
</tr>
<tr>
<td>Crowd</td>
<td>Interview, Telephone</td>
<td>Interview, Telephone</td>
</tr>
<tr>
<td>Single Speaker</td>
<td>Telephone</td>
<td>_</td>
</tr>
</tbody>
</table>

Figure 1 illustrates, for one leading system, the performance effects of the levels of added HVAC noise on 300 s interview test segments. This plot is on the extended data trials, producing smooth curves, and shows the considerably greater performance impact of 6 dB SNR compared with that of 15 dB SNR.

Figure 2 illustrates, for one leading system, the comparatively greater impact on performance of crowd noise than of HVAC noise at 6 dB SNR, relative to the no-noise condition.

Figure 3 similarly examines the effect of noise type for one leading system for 300 s telephone channel segments. Single speaker noise not surprisingly had greater impact than crowd noise. The telephone segments included some recorded with high environmental noise. The segments included for the environmental noise curve are those judged by LDC auditors.
to have been highest in environmental noise. (The auditors’ noise judgments often differed from those of the recorded subjects.) Note that the performance impact of environment noise is close to that for 6 dB SNR single speaker noise.

6. Duration

Figure 4 shows performance for telephone channel segments of 300 s, 100 s, and 30 s durations for one leading system. Since these are two-sided conversational segments, the actual amount of speech per segment generally averages half of these nominal durations.

Consistent with past experience, the largest difference is seen between the two shorter durations considered, with lesser improvement resulting from utilizing still longer durations. There is, however, considerable performance gain observed for 300 s compared to 100 s duration.

7. Training

To examine the effect of the quantity of training data, Figure 5 separates the trials involving telephone channel test segments according to whether the target speaker training consisted of single session or involved four or more sessions, for one leading system. The result is in the direction to be expected.

Only a few participants submitted alternative systems exploring the effect of the type of training data on performance. Figure 6 shows results on telephone test segments for one site with alternate systems that did training only on telephone or only on microphone speech, limiting the trials to those having target speakers with four or more training sessions, which helps avoid confounding with other factors. Except for the extremely low false alarm rate range, where limited trials (non-extended data) make results unstable, the advantage of matched channel types between training and test is apparent.

8. Channel

In the LDC collection protocol, speech data was collected from over a dozen different room microphones, but for SRE12 the trials were limited to data from four of these channels, designated 02, 04, 07, and 12. Figure 7 shows performance curves for interview trials limited to each of these four channels.

In the LDC collection protocol, speech data was collected from over a dozen different room microphones, but for SRE12 the trials were limited to data from four of these channels, designated 02, 04, 07, and 12. Figure 7 shows performance curves for interview trials limited to each of these four channels.

9. Sex

Past evaluations have shown some variation in performance by sex, with a frequent trend toward narrowly better performance for male speakers. Results for SRE12 showed little performance difference by sex except perhaps in the very low false alarm rate region. Figure 8 shows extended data curves for male and female speakers on telephone channel segments for one leading system.

Channels 02 and 04 were high quality microphones near the speaker, the former the lavalier worn by the subject, the latter placed on the table directly in front of him/her. Channels 07 and 12 were far-field microphones placed at some distance from the subject in the collection room. The performance effects of these different microphone types and placements are apparent.
SRE12 included more speech data and more trials than any of the prior NIST SRE’s, and this supported examination of a multiplicity of factors affecting performance. The evaluation was most notable for its use primarily of known target speakers (discussed elsewhere) and for its inclusion of test segments with multiple levels and types of additive noise. This was the first SRE to include additive noise, but this type of testing was also included in the NIST BEST interim assessment [6]. Since inclusion of such data does not require collection of additional speech, which is always expensive, and since systems have become increasingly capable of processing larger numbers of trials (in the millions or hundreds of millions) and of test segments (in the tens of thousands), future evaluations are likely to continue such systematic study of the effects of additive noise.

The other performance factors considered here have been studied in prior SRE’s and should remain of interest in the future.

The date of the next speaker evaluation has not yet been determined, but may be expected in the coming couple of years. It will include newly collected telephone data and may emphasize achieving performance independent of the location of the collection environment. It will, as always, be open to interested participants from all over willing to take part in accordance with the evaluation rules.

11. Disclaimer

These results are not to be construed or represented as endorsements of any participant’s system, methods, or commercial product, or as official findings on the part of NIST or the U.S. Government.

Certain commercial equipment, instruments, software, or materials may be identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the equipment, instruments, software or materials are necessarily the best available for the purpose.

12. References