NIST 2016 Speaker Recognition Evaluation Plan

August 4, 2016

1 Introduction

The 2016 speaker recognition evaluation (SRE16) is the next in an ongoing series of speaker recognition evaluations conducted by NIST since 1996. These evaluations serve (1) to support speaker recognition research by exploring promising new ideas in speaker recognition and developing advanced technology incorporating these ideas and (2) to measure and calibrate the performance of speaker recognition systems. They are intended to be of interest to all researchers working on the general problem of text independent speaker recognition. To this end the evaluation is designed to be simple, to focus on core technology issues, to be fully supported, and to be accessible to those wishing to participate.

The basic task in NIST's speaker recognition evaluations is speaker detection, i.e., to determine whether a specified target speaker is speaking during a given segment of speech. Like previous SREs, SRE16 focuses on telephone speech recorded over a variety of handset types. However, there are several differences with previous SREs:

- Target speaker data will not be distributed in advance like in SRE12
- Fixed training condition is introduced to allow better cross-system comparisons
- Test segments will have more duration variability than in previous evaluations
- The enrollment and test data were collected outside North America
- The evaluation will be conducted using same and different phone number trials

Participation in the evaluation is open to all who find the evaluation of interest and are able to comply with the evaluation rules set forth in this plan. There is no cost to participate, but participants must be represented at the evaluation workshop to be held in San Juan, Puerto Rico on December 11-12, 2016. Information about evaluation registration can be found on the SRE16 website[1].

2 Task Description

2.1 Task Definition

As stated in the Introduction, the task for SRE16 is *speaker detection*: given a segment of speech and the target speaker enrollment data, automatically determine whether the target speaker is speaking in the segment. A segment of speech (test segment) along with the enrollment speech segments(s) from a designated target speaker constitutes a *trial*. The system is required to process each trial independently and to output a log likelihood ratio (LLR) for that trial. The LLR is defined as follows:

$$LLR = ln\left(\frac{pdf(data|TargetHyp.)}{pdf(data|NonTargetHyp.)}\right) \tag{1}$$

[1] http://www.nist.gov/itl/iad/mig/sre16.cfm

2.2 Training Conditions

The training condition is defined as the amount of data/resources used to build a Speaker Recognition (SR) system. The task described above can be evaluated over a *fixed* (required) or *open* (optional) training condition.

- **Fixed** – The fixed training condition limits the system training to specific data sets. They are:
  - data provided from the new corpus collection
  - previous SRE data
  - Switchboard corpora that contain transcripts
  - Fisher corpora

  The LDC license lists the actual catalog numbers of these corpora. Participants can obtain the data from the Linguistic Data Consortium (LDC) after they have signed the LDC data license agreement. For the fixed training condition, only the specified speech data may be used for system training and development, to include all sub-systems (e.g., speech activity detection) and auxiliary systems used for automatic labels/processing (e.g., language recognition). Publicly available, non-speech audio and data (e.g., noise samples, impulse responses, filters) may be used and should be noted in the system description. Participation in this condition is required.

- **Open** – The open training condition removes the limitations of the fixed condition. In addition to the data listed in the fixed condition, participants can use other publicly available data. LDC will make selected data from the IARPA Babel Program to be used in the open training condition. Participation in this condition is optional but encouraged.

Sites are strongly encouraged to participate in both the fixed and open conditions to demonstrate the gains that can be achieved with unconstrained amounts of data.

2.3 Enrollment Conditions

The enrollment condition is defined as the number of speech segments provided to create a target speaker model. However, unlike previous SREs, gender labels will not be provided. There are two enrollment conditions for SRE16:

- **One-segment** – the system is given only one approximately 60 secs[2] of segment to build the model of the target speaker.
- **Three-segment** – the system is given three approximately 60 secs segments to build the model of the target speaker, all from the same phone number.

2.4 Test Conditions

- The test segments will be uniformly sampled ranging approximately from 10 secs to 60 secs. The test segments that are less than 9 secs will not be included in the primary metric calculation but will be scored for analysis of systems' behavior.
- Trials will be conducted with test segments from both same and different phone numbers as the enrollment segment(s).
- There will be no cross-sex trials.
- There will be no cross-language trials.

[2] as determined by SAD output

3 Performance Measurement

3.1 Primary Metric

A basic cost model is used to measure the speaker detection performance and is defined as a weighted sum of miss and false alarm error probabilities:

$$\begin{aligned} C_{Det}(C_{Miss}, C_{FalseAlarm}, P_{Target}) = & C_{Miss} \times P_{Target} \times P_{Miss|Target} + \\ & C_{FalseAlarm} \times (1 - P_{Target}) \times P_{FalseAlarm|NonTarget} \end{aligned} \tag{2}$$

where the parameters of the cost function are C_{Miss} (cost of a missed detection) and $C_{FalseAlarm}$ (cost of a spurious detection), and P_{Target} (a priori probability of the specified target speaker) and are defined to have the following values:

Parameter ID	C_{Miss}	$C_{FalseAlarm}$	P_{Target}
1	1	1	0.01
2	1	1	0.005

Table 1: SRE16 cost parameters

To improve the intuitive meaning of C_{Det}, it will be normalized by dividing it by $C_{Default}$, defined as the best cost that could be obtained without processing the input data (i.e., by either always accepting or always rejecting the segment speaker as matching the target speaker, whichever gives the lower cost):

$$C_{Norm} = \frac{C_{Det}}{C_{Default}} \tag{3}$$

where $C_{Default}$ is defined as:

$$C_{Default} = min \begin{cases} C_{Miss} \times P_{Target}, \\ C_{FalseAlarm} \times (1 - P_{Target}) \end{cases} \tag{4}$$

Substituting either set of parameter values from Table 1 into Equation 4 yields:

$$C_{Default} = C_{Miss} \times P_{Target} \tag{5}$$

Substituting C_{Det} and $C_{Default}$ in Equation 3 with Equations 2 and 5, respectively, along with some algebraic manipulations yields:

$$C_{Norm} = P_{Miss|Target} + \beta \times P_{FalseAlarm|NonTarget} \tag{6}$$

where β is defined as:

$$\beta = \frac{C_{FalseAlarm}}{C_{Miss}} \times \frac{1 - P_{Target}}{P_{Target}} \tag{7}$$

Actual detection costs will be computed from the trial scores by applying detection thresholds of $\log(\beta)$ for the two values of β, with β_1 for $P_{Target_1} = 0.01$ and β_2 for $P_{Target_2} = 0.005$. Thus, the primary cost measure for SRE16 is defined as:

$$C_{Primary} = \frac{C_{Norm_{\beta_1}} + C_{Norm_{\beta 2}}}{2} \tag{8}$$

The evaluation data will be divided into 16 partitions. Each partition is defined as a combination of enrollment (1-segment or 3-segment), language (Tagalog or Cantonese), sex (Male or Female), and phone number match (same or different). $C_{Primary}$ will be calculated for each partition, and the final results is the average of all the partitions' $C_{Primary}$'s.

Also, a minimum detection cost will be computed by using the detection thresholds that minimize the detection cost. Note that for minimum cost calculations, the counts for each condition set will be equalized before pooling and cost calculation (i.e., minimum cost will be computed using a single threshold not one per condition set).

NIST will make available the script that calculates the primary metric.

3.2 Alternative Metric

In addition to the primary metric, an alternative, information theoretic measure may be computed that considers how well all scores represent the likelihood ratio and that penalizes for errors in score calibration. This performance measure is defined as:

$$C_{llr} = \frac{1}{2 \times \log(2)} \times \left(\frac{\sum \log(1 + \frac{1}{s})}{N_{TT}} + \frac{\sum \log(1 + s)}{N_{NT}} \right) \tag{9}$$

where the first summation is over all target trials N_{TT}, the second is over all non-target trials N_{NT}, and s represents a trial's likelihood ratio[3].

4 Data Description

The data collected by the LDC as part of the Call My Net Speech Collection to support speaker recognition research will be used to compile the SRE16 test set, development set, and part of the training set[4].

The data are composed of telephone conversations collected outside North America, spoken in Tagalog and Cantonese (referred to as the *major* language) and Cebuano and Mandarin (referred to as the *minor* languages). The development set described below will contain data from both the major and minor languages, while the test set will be contain data from the two major languages. Recruited speakers (called *claque* speakers) made multiple calls to people in their social network (e.g., family, friends). Claque speakers were encouraged to use different telephone instruments (e.g., cell phone, landline) in a variety of settings (e.g., noisy cafe, quiet office) for their initiated calls and were instructed to talk for 10 minutes on a topic of their choosing.

All segments will be encoded as a-law sampled at 8kHz in SPHERE formatted files. The development and test sets will be distributed by NIST via Amazon Web Services (AWS).

4.1 Data Organization

The development and test sets follow a similar directory structure:

```
<base_directory>/
        README.txt
        data/
            enrollment/
            test/
            unlabeled/ (in training set only)
        docs/
        metadata/ (in development set only)
```

[3] The reasons for choosing this cost function, and its possible interpretations, are described in detail in the paper "Application-independent evaluation of speaker detection" in Computer Speech & Language, volume 20, issues 2-3, April-July 2006, pages 230-275, by Niko Brummer and Johan du Preez.

[4] The entire training set also includes previous SRE corpora, Switchboard, and Fisher corpora. See Section 4.4.

4.2 Trial File

The trial file, named `sre16_{dev|eval}_trials.tsv` and located in the `docs` directory, is composed of a header and a set of records where each record describes a given trial. Each record is a single line containing three fields separated by a tab character and in the following format:

```
modelid<TAB>segment<TAB>side<NEWLINE>
```

where
`modelid` - The enrollment identifier
`segment` - The test segment identifier
`side` - The channel[5]

For example:

```
modelid segment side
1001_sre16 dtadhlw_sre16 a
1001_sre16 dtaekaz_sre16 a
1001_sre16 dtaekbb_sre16 a
```

4.3 Development Set

Participants in the SRE16 evaluation will receive data for development experiments that will mirror the evaluation conditions. The development data will be drawn from the minor languages and will include:

- 20 speakers, 10 each from Cebuano and Mandarin
- 10 calls per speaker
- Associated metadata which will be listed in the following files located in the `metadata` directory as outlined in section 4.1.
  - calls.tsv - information about the calls (e.g., conversations)
  - call_sides.tsv - information about the call sides
  - languages.tsv - information about the languages
  - subjects.tsv - information about the speakers

The development data may be used for any purpose.

4.4 Training Set

Section 2.2 describes the two training conditions: Fixed (required) and Open (optional). Participants in the SRE16 evaluation will receive a common set of data resources for training for the fixed training condition. An unlabeled (i.e., no speaker id, gender, language, or phone number information) set of approximately 2200 calls from the Call My Net collection will be made available divided into sets from the minor and major languages. In addition participants will receive data from all previous SRE corpora as well as Switchboard corpora that contain transcripts and the Fisher corpus with transcripts. To obtain this set, participants must sign the LDC data license agreement which outlines the terms of the data usage.

Additionally, LDC will be releasing selected data resources from the IARPA Babel Program for use in the open training condition. All training sets will be available directly from the LDC[6].

Participants are encouraged to submit results for the contrastive open training condition to demonstrate the value of additional data.

[5] SRE16 segments will be single channel so this field is always ”a”

[6] http://www.ldc.upenn.edu

5 Evaluation Rules and Requirements

SRE16 is conducted as an open evaluation where the test data is sent to the participants who process the data locally and submit the output of their systems to NIST for scoring. As such, the participants have agreed to process the data in accordance with the following rules:

- The participants agree to abide by the terms guiding the training conditions (fixed or open).
- The participants agree to process at least the fixed training condition.
- The participants agree to process each trial independently. That is, each decision for a trial is to be based only upon the specified test segment and target speaker enrollment data. The use of information about other test segments and/or other target speaker data is not allowed.
- The participants agree not to probe the enrollment or test segments via manual/human means such as listening to the data or producing the transcript of the speech.
- The participants agree not to produce manual/human annotations of the unlabeled training data, such as employing a service like Amazon's Mechanical Turk. Informal listening and spectral analysis of subsets of the audio are acceptable.
- The participants are allowed to use any automatically derived information for training, development, enrollment, test segments, provided that the automatic system used conforms to the training data condition (fixed or open) for which it is used.
- The participants are allowed to use information available in the SPHERE header.
- The participants can submit up to three systems per training condition. Bug-fix does not count toward this limit.

In addition to the above data processing rules, participants agree to comply with the following general requirements:

- The participants agree to have one or more representatives at the evaluation workshop to present a meaningful description of their system(s). Evaluation participants failing to do so will be excluded from future evaluation participation.
- The participants agree to the guidelines governing the publication of the results:
  - Participants are free to publish results for their own system but must not publicly compare their results with other participants (ranking, score differences, etc.) without explicit written consent from the other participants.
  - While participants may report their own results, participants may not make advertising claims about winning the evaluation or claim NIST endorsement of their system(s). The following language in the U.S. Code of Federal Regulations (15 C.F.R. § 200.113) shall be respected[7]: *NIST does not approve, recommend, or endorse any proprietary product or proprietary material. No reference shall be made to NIST, or to reports or results furnished by NIST in any advertising or sales promotion which would indicate or imply that NIST approves, recommends, or endorses any proprietary product or proprietary material, or which has as its purpose an intent to cause directly or indirectly the advertised product to be used or purchased because of NIST test reports or results.*
  - At the conclusion of the evaluation NIST generates a report summarizing the system results for conditions of interest, but these results/charts do not contain the participant names of the systems involved. Participants may publish or otherwise disseminate these charts, unaltered and with appropriate reference to their source.

[7] See http://www.ecfr.gov/cgi-bin/ECFR?page=browse

- The report that NIST creates should not be construed or represented as endorsements for any participant's system or commercial product, or as official findings on the part of NIST or the U.S. Government.

6 Evaluation Protocol

To facilitate information exchange between the participants and NIST, all evaluation activities are conducted over a web-interface.

6.1 Evaluation Account

Participants must sign up for an evaluation account where they can perform various activities such as registering for the evaluation, signing the data license agreement, uploading the submission and system description. To sign up for an evaluation account, go to https://sre.nist.gov. The password must be at least 12 characters long and must contain a mix of upper and lowercase letters, numbers, and symbols. After the evaluation account is confirmed, the participant is asked to join a site or create one if it does not exist. The participant is also asked to associate his site to a team or create one if it does not exist. This allows multiple members with their individual accounts to perform activities on behalf of their site and/or team (e.g., make a submission) in addition to performing their own activities (e.g., requesting workshop invitation letter).

- A site is defined as a single organization (e.g., NIST)
- A team is defined as a group of organizations collaborating on a task (e.g., Team1 consisting of NIST and LDC)
- A participant is defined as a member or representative of a site who takes part in the evaluation (e.g., John Doe)

6.2 Evaluation Registration

One participant from a site must formally register his site to participate in the evaluation by agreeing to the terms of participation. For more information about the terms of participation, see Section 5.

6.3 Data License Agreement

One participant from each site must sign the LDC data license agreement to obtain the training data for the fixed training condition and Babel data for the open training condition.

6.4 Submission Requirements

Each team must participate in the fixed training condition. Teams are encouraged to participate in the open training condition to demonstrate the gains that can be achieved with unconstrained amounts of data. For each training condition, the team can submit up to three systems and must designate one as the *primary* system that NIST uses for cross-team comparisons. There should be one output file for each training condition per system.

Each team is required to submit a system description at the designated time (see Section 7). The evaluation results are given only after the system description is received.

6.4.1 System Output Format

The system output file is composed of a header and a set of records where each record contains a trial given in the trial file (see Section 4.2) and a log likelihood ratio output by the system for the trial. The order of the trials in the system output file must follow the same order as the trial list. Each record is a single line containing 4 fields separated by tab character in the following format:

```
modelid<TAB>segment<TAB>side<TAB>llr<NEWLINE>
```

where
`modelid` - The enrollment identifier
`segment` - The test segment identifier
`side` - The channel (always "a" for SRE16 since the data is single channel)
`llr` - The log likelihood ratio

For example:

```
modelid segment side llr
1001_sre16 dtadhlw_sre16 a 0.79402
1001_sre16 dtaekaz_sre16 a 0.24256
1001_sre16 dtaekbb_sre16 a 0.01038
```

There should be one output file for each training condition for each system. NIST will make available the script that validates the system output.

6.4.2 System Description Format

Each team is required to submit a system description. The system description must include a brief description of the systems/algorithms used to produce the results and a timing report. The timing report describes the CPU execution time that is required to process the test set as if running on a single CPU and as a multiple of real-time for the data processed. The timing report should identify the time for creating models from the enrollment data and the time needed for processing the test segments. The timing report should include the CPU(s) utilized and the amounts of memory used. The system description should follow the IEEE conference proceeding template. A copy of the template is available on the SRE16 website.

7 Schedule

Milestone	Date
Evaluation plan published	March 2016
Registration period	April 19 - September 13, 2016
Training data available	May, 2016
Evaluation data available to participants	September 20, 2016
System output due to NIST	October 11, 2016
Preliminary results released	October 25, 2016
Post evaluation workshop co-located with SLT in San Juan, Puerto Rico	December 11-12, 2016