I4U submission to NIST SRE 2012:
A large-scale collaborative effort for noise-robust speaker verification


1Radboud University Nijmegen, The Netherlands, 2Institute for Infocomm Research, Singapore, 3University of Eastern Finland, Finland 4CRSS, University of Texas at Dallas, USA, 5ValidSoft Ltd, London, UK, 6LIA, University of Avignon, France 7Idiap Research Institute, Switzerland, 8Swansean University, UK, 9University of New South Wales, Australia

Abstract

I4U is a joint entry of nine research Institutes and Universities across four continents to NIST SRE 2012. It started with a brief discussion during the Odyssey 2012 workshop in Singapore. An online discussion group was soon set up, providing a discussion platform for different issues surrounding NIST SRE’12. Noisy test segments, uneven multi-session training, variable enrollment duration, and the issue of open-set identification were actively discussed leading to various solutions integrated to the I4U submission. The joint submission and several of its 17 subsystems were among top-performing systems. We summarize the lessons learnt from this large-scale effort.

Index Terms: Speaker Verification, NIST SRE 2012, I4U, i-vector

1. Introduction

The I4U submission to National Institute of Standards and Technology (NIST) speaker recognition evaluation 2012 (SRE’12) [1] is a result of active exchange of information between the coalition participants across nine institutions. The name of the institutes and corresponding system identifiers are provided in Table 1. The submitted results are based on the fusion of multiple classifiers. The optimization of the component classifiers and the fusion device were done with development sets jointly designed within the I4U coalition with multiple design iterations, refinement of noise adding protocol and various other details. Different from previous SREs, the task of SRE’12 involves:

Handling noisy test segments: This required speech enhancement algorithms and employing mixed training or parallel model combination techniques.

Imbalanced multi-session training: There are tens of segments available for training some speaker models while only a single segment for some other speakers.

Open-set identification: SRE’12 evaluation protocol allows the use of knowledge of all target speakers in each detection trials which resulted in utilizing compound log-likelihood ratio.

Table 1: I4U Coalition and assigned system indexes

<table>
<thead>
<tr>
<th>Site</th>
<th>System index</th>
</tr>
</thead>
<tbody>
<tr>
<td>ValidSoft Ltd (VLD)</td>
<td>Sys1</td>
</tr>
<tr>
<td>Swansea University (UWS)</td>
<td>Sys2</td>
</tr>
<tr>
<td>University of Avignon (LIA)</td>
<td>Sys3</td>
</tr>
<tr>
<td>Radboud University Nijmegen (RUN)</td>
<td>Sys4</td>
</tr>
<tr>
<td>University of Texas at Dallas (CRSS)</td>
<td>Sys5–10</td>
</tr>
<tr>
<td>University of Eastern Finland (UEF)</td>
<td>Sys11</td>
</tr>
<tr>
<td>Institute for Infocomm Research (IIR)</td>
<td>Sys12–16</td>
</tr>
<tr>
<td>Idiap Research Institute (IDIAP)</td>
<td>Sys17</td>
</tr>
</tbody>
</table>

This paper is organized as follows: In Section 2, we present the strategies taken to make a development set coping with SRE’12 new conditions. Details of the submitted systems and the component classifiers, together with the strategies to deal with the new challenges listed above are described in Section 3. One of the motivations underlying the I4U coalition is to experiment with the fusion of large numbers of sub-systems. Results for the individual and the fused system are presented in Section 4.

2. Development sets

The development sets were generated to help I4U team members in developing their speaker recognition systems considering the special conditions in SRE’12 including multiple segment training for a speaker 1. All the members of I4U coalition helped in refining the lists with respect to detecting empty or otherwise problematic segments with conflicts in gender and speaker PIN (there are issues with pre-SRE’12 lists like multiple-genders or wrong genders for some speakers). The latest lists from NIST were utilized and speech segments for all 1918 target speakers were fetched from SRE’06, SRE’08 and SRE’10 corpora and corresponding meta-data were extracted. To be able to assess both the recognition systems’ generalization and calibration performance, separate development (DEV) and evaluation (EVAL) sets were created. The number of segments, speakers and trials for each set are given in Table 2. In designing these sets, the following criteria were considered:

- Test segments are disjoint for DEV-test and EVAL-test.
- Most of the train segments in DEV-train are added to EVAL-train. The number of train segments in EVAL-train is almost

---

This work was partly supported by Academy of Finland (proj. no 253120 and 132129), Swiss National Science Foundation under the LOBI project, contract no. SNSF-235 and European Community’s Seventh Framework Program (FP7/2007-2013) under grant agreement no. 238803.

1 The lists are available via http://cls.ru.nl/~saeidi/file_library/I4U.tgz
twice the number of segments in DEV-train. This design choice is made to evaluate the systems performance under the condition that speaker and channel spaces are already trained but the number of enrollment segments for target speaker modeling has increased.

- The segments from train to test have all different LDC-IDs to avoid testing against same session from training.
- Two disjoint sets of speakers from SRE’06 data that do not appear in SRE’12 are added to DEV-test and EVAL-test to form unknown non-target trials.
- For those speakers having telephone and microphone data, both types of channels were included in the train set so that systems could benefit from having different channels in training.
- Considering noisy segments inclusion in NIST SRE’12, for every original NIST segment, two noisy versions were generated. Noise adding was carried out using FaNT-3. We have used ten noise segments for each HVAC (heating, ventilation and air-conditioning) and crowd noise type. Noise signals used to contaminate the speech segments were different from train to test and from DEV to EVAL. Noises are added at two SNR-levels 6dB and 15dB. The mean measured SNR-levels were 40dB, 15dB and 10dB for original, 15dB and 6dB segments, respectively. Since there are two noisy versions of each clean segment being utilized in DEV and EVAL sets, the performance of the developed systems are optimized to perform well under noisy condition rather than clean (“not altered”) condition.

### 3. Recognition systems

The systems developed in the I4U coalition were based on state-of-the-art: 1) i-vector system [11] with probabilistic linear discriminant analysis (PLDA) [12] modeling, or 2) Gaussian supervector representation and joint factor analysis (JFA) [13, 14], or support vector machine (SVM) modeling. All 16 kHz audio data were down-sampled to 8 kHz to match to the existing 8 kHz background data. Energy-based speech activity detection (SAD) was applied to telephone segments, while for interview segments a dual-channel SAD is employed. The automatic speech recognition (ASR) transcripts from NIST for interview segments in SRE’08 and SRE’10 were used to refine the SAD labels. All of the systems are gender-dependent. The components and data usage of sub-systems are presented in Tables 3, 4 and 5 for features, transform and classifier, respectively.

#### Sys1: Validsoft’s i-vector system uses spectral subtraction to enhance energy profile for SAD. Test i-vectors are scored against all target segment i-vectors followed by score averaging.

#### Sys2: Swansea’s i-vectors are normalized with eigen-factors radial (EFR) method [15] utilizing total covariance matrix of the background data. LDA-reduced 200-dimensional i-vectors are averaged for each target speaker and used with Mahalanobis scoring.

#### Sys3: LIA’s system uses two fused sub-systems. The first uses LDA reduction preceded by iterative i-vector normalization according to the covariance matrix [15] and two-covariance scoring; the second uses PLDA preceded by spherical nuisance normalization within-class covariance matrix [15]. Score is computing as a) the average score of the test i-vector against all target i-vectors and b) an equal-weight combination of these scores according to multiple PLDA subspace dimensions (from 50 to 400 in steps of 50).

#### Sys4: RUN’s i-vector PLDA system uses dynamic noise suppression within a Wiener filter applied both for speech enhancement and SAD. Noise estimation uses improved minima controlled recursive averaging (IMCRA) [5, 19] which averages the previous estimate of the noise power spectra and has proven robust against input SNR and different noise types due to rapid noise tracking. The noise power spectral density estimate is used for decision-directed a-priori SAD estimation, which further defines a Wiener filter applied for magnitude enhancement.

#### Sys5-10: The CRSS’s i-vector systems use combinations of two different front-ends and three back-ends [20, 21]. Gaussianized cosine-distance scoring (GCDS) and a discriminative back-end using L2-regularized logistic regression (using LIBLINEAR [18]) are used. The enrollment i-vectors are averaged and then Gaussianized using mean and variance of devset. LDA dimensionality reduction and cosine scoring are used.

#### Sys11: UCM contributed the overall fusion component for I4U [22, 23] and developed a robust utterance-adaptive SAD [10] where 16-component speech and non-speech codebooks are trained from 12 MFCCs including c0. Training labels available at [http://cs.uef.fi/pages/kiinnu/VQVAD/VQVAD.zip](http://cs.uef.fi/pages/kiinnu/VQVAD/VQVAD.zip)
are obtained from reliable frames with the help of aggressive spectral oversubtraction. The recognizer is a standard i-vector PLDA system and, unlike most of the other I4U system, does not use multicondition training.

Sys12 by I2R whitens the first-order sufficient statistics using UBM covariances, which speeds up estimation of the posterior distribution during the total variability matrix (T-matrix) training and i-vector extraction [24]. T-matrix estimation uses two subspaces, $T_{tel}$ and $T_{mic}$, where $T_{tel}$ is trained from telephone data and $T_{mic}$ from microphone data following decoupled method on [25]. This enables easy control of the dimensionality of the subspaces in $T = [T_{tel}; T_{mic}]$ and avoids the problem of data type imbalance encountered when all data are pooled for T-matrix training in one go. For details of the PLDA implementation, refer to [26].

Sys13 by I2R is a GMM supervector system with KL divergence kernel [27]. Utterance GMM is obtained via MAP adaptation of the UBM means that are concatenated and normalized by the UBM standard deviation and square root of the mixture weights. Nuisance attribute projection (NAP) [28] and tz-norm are applied for channel and score normalization, respectively.

Sys14 is an anti-model variant of Sys13. The use of other target speakers is allowed in SRE’12 which leads to an open-set identification problem. The anti-model approach of [29] is adopted for increased discrimination between target and unseen non-targets. SVM for each target speaker is trained using the sub-vectors of the other target speakers as the SVM background together with additional data drawn from SRE’04 for the unseen speakers.

Sys15 is a Bhattacharyya-kernel GMM-SVM system with data-dependent relevance factor [30, 31] and zt-norm. Sys16, in turn, uses joint factor analysis (JFA) implementation for I2R’s SRE’10 submission [32]. It is composed of 300 speaker factors, 200 channel factors (100 for telephone, 50 for microphone, 50 for interview), and full rank diagonal matrix. For eigenchannel training, the tel, mic and interview channels were separately trained and concatenated into an eigenchannel matrix. Enrollment and scoring (with zt-norm) are as in Sys15.

Sys17: IDIAP’s system is a single classifier with inter-session variability (ISV) modeling technique [16]. It is implemented using Bob’s, an open-source signal processing and machine learning toolbox. ISV is similar to JFA with linear scoring approximation [33] but with merged eigen-voice and -channel spaces. Scores are normalized using zt-norm.

### 4. System performance

We analyze and compare system performance on the core task of NIST SRE’12 using the equal error rate (EER) and primary cost. The notion of EER is commonly known. What is new in SRE’12 is the use of the so-called primary cost $C_{primary}$, defined as the average cost at two specific points on the DET curve. At either of these points, the detection cost function (DCF) is defined in normalized form (such that the maximum value is one), as follows:

$$C_{Norm}(\theta) = P_{false} + \frac{1}{2} \left[ P_{false} + P_{false|unknown} \right]$$

where $P_{false}$ is the false reject rate and $P_{false|unknown}$ is the false acceptance rate. The normalized cost is used to assess the performance of the different systems.
Here, $P_{\text{tar}}$ is the a priori probability that a trial is a target trial, while $P_{\text{miss}}(\theta)$ and $P_{\text{fa}}(\theta)$ are, respectively, the probability of miss and false alarm at threshold $\theta$. Notice that $P_{\text{fa}}(\theta)$ consists of two components computed separately from the known and unknown non-target trials. Now, let $\theta_A$ be the threshold which gives the $C_{\text{norm}}(\theta_A)$ with $P_{\text{tar}} = 0.01$ and $\theta_B$ be the threshold which gives the $C_{\text{norm}}(\theta_B)$ with $P_{\text{tar}} = 0.001$, the primary detection cost is defined as the average cost between the points on the detection error trade-off (DET) curve, as follows

$$C_{\text{primary}} = \frac{C_{\text{norm}}(\theta_A) + C_{\text{norm}}(\theta_B)}{2}$$

Table 6 shows the absolute performance of all 17 systems and their fusion for common conditions 1 and 3 as defined in SRE’12. One obvious point to note here is that, the PLDA i-vector systems give consistently better performance in terms of EER and minimum $C_{\text{primary}}$ when the test signal is collected over clean (CC1) and noisy (CC3) interview sessions. It is also obvious that, the GMM-SVM (Sys 13, 14, and 15) and JFA (Sys 16) gave equally good performance compared to, and for some instances better than i-vector based systems.

![Figure 1: Analysis of excluding one system at a time in fusion using Focal and employing compound log-likelihood ratio [34] for $P_{\text{known}} = 0.5$. Using the full ensemble of classifiers results in actual $C_{\text{primary}}$ of 0.3959 and 0.2836 for first two common conditions (CC1 and CC2) respectively in SRE’12 for the pooled scores of males and females. A positive relative change indicates increased actual $C_{\text{primary}}$ by excluding a system in fusion resulting in fusion performance drop. Systems number 2 and 3 are not considered for this analysis.](image)

5. Conclusion

This paper provides an overview of fusion of 17 systems submitted to NIST SRE’12 by different sites in I4U coalition. The collaboration of over 30 researchers within the coalition benefited all the sites in preparing robust speaker recognition systems. It is hard to compare the individual subsystems and determine the strengths of each system but in a very general prospective, the systems that utilized more recent features and employ speech enhancement in the front-end were more successful. Averaging the enrollment i-vectors gave about the same performance as averaging the scores of i-vectors. Discriminative training schemes, such as SVMs, using a proper distance kernel on Gaussian supervisor representation was found to outperform generative i-vector representation with PLDA classification. The new paradigm shift in NIST SRE’12 is expected to emphasize the discriminative training in modeling and even i-vector representation.

---

7 http://niko.brummer.googlepages.com/focal
6. References


