

# The I4U Mega Fusion and Collaboration for NIST Speaker Recognition Evaluation 2016

K A Lee, V Hautamäki, T Kinnunen, A Larcher, C Zhang, A Nautsch, T Stafylakis, G Liu, M Rouvier, W Rao, et al.

▶ **To cite this version:**

K A Lee, V Hautamäki, T Kinnunen, A Larcher, C Zhang, et al.. The I4U Mega Fusion and Collaboration for NIST Speaker Recognition Evaluation 2016. Annual Conference of the International Association of Speech Communication (Interspeech), Aug 2017, Stockholm, Sweden. hal-01927561

**HAL Id: hal-01927561**
**https://hal.science/hal-01927561**

Submitted on 19 Nov 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# The I4U Mega Fusion and Collaboration for NIST Speaker Recognition Evaluation 2016

*K. A. Lee*[1], *V. Hautamäki*[2], *T. Kinnunen*[2], *A. Larcher*[3], *C. Zhang*[4], *A. Nautsch*[5], *T. Stafylakis*[6], *G. Liu*[7], *M. Rouvier*[8], *W. Rao*[9], *F. Alegre*[10], *J. Ma*[11], *M. W. Mak*[12], *A. K. Sarkar*[13], *H. Delgado*[14], *R. Saeidi*[15], *H. Aronowitz*[16], *A. Sizov*[1], *H. Sun*[1], *T. H. Nguyen*[1], *G. Wang*[1,†], *B. Ma*[1], *V. Vestman*[2], *M. Sahidullah*[2], *M. Halonen*[2], *A. Kanervisto*[2], *G. Le Lan*[3], *F. Bahmaninezhad*[4], *S. Isadskiy*[5], *C. Rathgeb*[5], *C. Busch*[5], *G. Tzimiropoulos*[6], *Q. Qian*, *Z. Wang*[7], *Q. Zhao*[7], *T. Wang*[7], *H. Li*[7], *J. Xue*[7], *S. Zhu*[7], *R. Jin*[7], *T. Zhao*[7], *P.-M. Bousquet*[8], *M. Ajili*[8], *W. B. Kheder*[8], *D. Matrouf*[8], *Z. H. Lim*[9], *C. Xu*[9], *H. Xu*[9], *X. Xiao*[9], *E. S. Chng*[9], *B. Fauve*[10], *K. Sriskandaraja*[11], *V. Sethu*[11], *W. W. Lin*[12], *D. A. L. Thomsen*[13], *Z.-H. Tan*[13], *M. Todisco*[14], *N. Evans*[14], *H. Li*[1], *J. H. L. Hansen*[4], *J.-F. Bonastre*[8], *E. Ambikairajah*[11]

[1]Institute for Infocomm Research (I$^2$R), A$^\star$STAR, Singapore, [2]University of Eastern Finland, Finland,
[3]LIUM, Universite du Maine, France, [4]CRSS, University of Texas at Dallas, USA,
[5]Hochschule Darmstadt, Germany, [6]University of Nottingham, UK, [7]Alibaba Inc., USA,
[8]LIA, University of Avignon, France, [9]Nanyang Technological University, Singapore,
[10]ValidSoft, UK, [11]University of New South Wales, Australia,
[12]The Hong Kong Polytechnic University, Hong Kong SAR, [13]Aalborg University, Denmark,
[14]EURECOM, France, [15]Aalto University, Finland, [16]IBM Research, Israel

## Abstract

The 2016 *speaker recognition evaluation* (SRE'16) is the latest edition in the series of benchmarking events conducted by the National Institute of Standards and Technology (NIST). I4U is a joint entry to SRE'16 as the result from the collaboration and active exchange of information among researchers from sixteen **I**nstitutes and **U**niversities across **4** continents. The joint submission and several of its 32 sub-systems were among top-performing systems. A lot of efforts have been devoted to two major challenges, namely, unlabeled training data and dataset shift from *Switchboard–Mixer* to the new *Call My Net* dataset. This paper summarizes the lessons learned, presents our shared view from the sixteen research groups on recent advances, major paradigm shift, and common tool chain used in speaker recognition as we have witnessed in SRE'16. More importantly, we look into the intriguing question of fusing a large ensemble of sub-systems and the potential benefit of large-scale collaboration.

**Index Terms**: speaker recognition evaluation, fusion, benchmark, Call My Net

## 1. Introduction

The speaker recognition evaluation (SRE) benchmark regularly conducted by the National Institute of Standards and Technology (NIST) has been a major driving force advancing speaker recognition technology. Since the first SRE'96 [1], the NIST evaluations have been focusing on speaker verification: given a segment of speech, decide whether a specified target speaker is speaking in that segment. Apart from the large datasets made available by the organizer, NIST SREs have been driving the research directions in speaker recognition by specifying the performance metrics and evaluation protocol – *conversational* versus *interview* speaking style in SRE'08 and SRE'10, *multiple* versus *single* session enrolment in SRE'12 [2], and more recently *fixed* versus *open* training conditions in SRE'16 [3], just to name a few. SRE'16 is different from the prior SREs in certain key aspects. For the first time, non-English evaluation data collected outside North America was used in SRE'16. In particular, the evaluation data is in Cantonese and Tagalog while the development set is in Mandarin and Cebuano. One more difference from previous editions is that SRE'16 explores the use of unlabeled dataset, which might hold the key to cope with the language mismatch and dataset shift problem [4].

Aside from a joint submission, the I4U consortium was formed with a common vision to promote and facilitate active exchange of information and experience toward SRE'16. Following the success of I4U'12 [5], the I4U consortium for SRE'16 is a collaboration of 62 researchers from 16 research **I**nstitutes and **U**niversities across **4** continents. The names of the organization and corresponding system identifiers are provided in Table 1. The collaboration started off with the first I4U meeting conducted via teleconference in early May, 2016. This was followed by regular bi-weekly and weekly meetings toward the end of SRE'16. An online group was also set up, providing a discussion platform across various issues surrounding NIST SRE'16. In particular, test segment variability, domain adaptation for language and channel shift, uncertainty propagation, score normalization, session compensation, and various issues concerning score calibration (quality measure, supervised versus unsupervised) have been actively discussed. Solutions were put in place as part of the I4U submission.

The submitted results were based on fusion of multiple classifiers from a pool of 32 sub-systems contributed by I4U members. The availability of such a large ensemble of sub-systems allows us to look into classifier selection and fusion strategies, and more importantly develop a strategy for *megafusion* [1] as detailed in Section 4. Listed in Table 1 are the 17 best sub-systems selected from each site to form the I4U primary submission.

The paper is organized as follows. Section 2 presents an overview of SRE'16 datasets and our strategies to cope with new conditions. Section 3 highlights some important aspects of the submitted system and the component classifiers. Results of individual sub-systems and their fusions are presented in Section 4. Section 5 concludes the paper.

[1]Thanks to Doug Reynolds of MITLL who coined the term 'megafusion' during the SRE'16 workshop, which has inspired the title of this paper. † Guangsen Wang is now with Tencent AI Lab, Shenzhen, China.

Table 1: *Component classifiers and their assigned system indexes used for I4U primary submission.*

| Sys | Feature and classifier | Site |
|---|---|---|
| 1 | PNCC IV-PLDA + GMM-UBM | AAU |
| 2 | MFCC IV-SVM | Alibaba Inc |
| 3 | MFCC IV-PLDA | CRSS, UTD |
| 4 | MFCC IV-PLDA | CRSS, UTD |
| 5 | MFCC IV-SVDA-PLDA | CRSS, UTD |
| 6 | ICMC IV-PLDA | EURECOM |
| 7 | MFCC IV-PLDA + Quality Meas Func | HDA |
| 8 | MFCC IV-PLDA | $I^2R$ |
| 9 | Tandem DNN IV-PLDA | $I^2R$ |
| 10 | MFCC DNN I-vector | LIUM |
| 11 | Tandem IV-PLDA | NTU |
| 12 | MFCC IV-PLDA | HK Poly U |
| 13 | MFCC IV-PLDA | UEF |
| 14 | MFCC IV-PLDA | EET, UNSW |
| 15 | PLP IV-PLDA | ValidSoft |
| 16 | MFCC DNN IV-PLDA | Nottingham |
| 17 | MFCC IV-PLDA | LIA |

## 2. Train, development, and test sets

To build and evaluate a speaker recognition system, three disjoint datasets are required for *training*, *development*, and *test*, respectively. In the context of NIST SREs, the training and development sets are usually released at a much earlier date than the test set (a.k.a the *evaluation* set), which the participants have to process and submit their system outputs (in the form of log-likelihood scores) in a limited period of time.

Similar to previous evaluations, the emphasis of SRE'16 has been on conversational telephone speech recorded from the public telephone system (e.g., cell phone, landline). Table 2 shows the list of corpora we used in I4U for the core task of SRE'16. In particular, the training set constitutes data used to train the *universal background model* (UBM) [6], *deep neural network* (DNN) to predict senone target [7, 8, 9], i-vector extractor [10], linear discriminant analysis (LDA), probabilistic LDA (PLDA) [11, 12, 13] and other parametric models found in most state-of-the-art systems, as in those top performance systems with I4U. It is worth mentioning that *Switchboard 1* (Swb 1) and the two *Fisher* corpora come with transcription which is required in training DNN [14, 15], or *splice time delay DNN* (TDNN) [16], to produce frame-level senone label and to extract deep bottleneck features [9, 17]. The *Fisher* corpora were also used for UBM training in some of I4U systems. Very specific to I4U was that we set aside SRE'10 and SRE'12 data for development and experiment purposes instead of including them in the training set. This setting allows us to validate the results across multiple validation and test sets.

Different from previous evaluations, the test data of SRE'16 consists of speech utterances spoken in Tagalog and Mandarin instead of English in previous evaluations. Also, the test segments have varying duration ranging from 10 to 60 seconds. These changes introduce two new challenges to SRE'16:

i. *Dataset shift* between the training and test sets due to language mismatch and differences in data collection infrastructure (e.g., telephone network, front-end devices).

ii. *Test duration variability* where test duration is uniformly distributed from 10 to 60 seconds.

The SRE'16 test data is a subset of the on-going *Call My Net* speech collection by the Linguistic Data Consortium (LDC). The changes from the Switchboard and Mixer datasets used in the previous SREs to the new *Call My Net* is cast as a dataset shift problem. In this regard, we found the *inter dataset variability compensation* (IDVC) [4] to be extremely effective in dealing with this problem. On the other hand, the duration variability was accounted for with the use uncertainty propagation [19] and variance compensated length-norm [20] for i-vector PLDA system. Most sub-systems developed in I4U were equipped with the dataset and duration compensation techniques mentioned above.

Also shown in Table 2 is the development set consisting of speech utterances spoken in Cebuano and Cantonese drawn from the same *Call My Net* collection as the test set. A small *unlabeled* set in Tagalog, Mandarin, Cebuano, and Cantonese was made available with an intention to bridge the gap between the development and test sets. Numerous efforts and discussion among I4U members have focused on the use of the unlabeled set for score normalization and calibration.

Table 2: *List of telephone speech copora partitioned into training, development, and test sets used in I4U for SRE'16.*

| Partition | Corpus | Language |
|---|---|---|
| Training | SRE'04-05-06-08<br>Swb-2 Phase II & III<br>Swb-Cell Part 1 & 2 | English |
| | Fisher 1 & 2<br>Swb-1 Release 2 | English<br>(with transcript) |
| | CallMyNet-Unlabeled | Cebuano, Mandarin<br>Tagalog, Cantonese |
| Development | CallMyNet-Dev | Cebuano, Mandarin |
| | SRE'10, SRE'12 | English |
| Test | CallMyNet-Test | Tagalog, Cantonese |

## 3. Recognition systems

Table 3 shows the key features of 17 systems used to form the I4U'16 primary submission. At the core of all the sub-systems listed in Table 3 is the i-vector approach [10], which represents the current mainstream technique in text-independent speaker recognition. At the i-vector extraction stage, we have sub-systems that use either GMM or DNN posteriors [7, 8] for frame alignment (i.e., the role of the UBM). Except for Sys2 that uses *support vector machine* (SVM), PLDA was used to handle session variability and as the scoring back-end in all other sub-systems. In addition, a rich set of acoustic feature extraction including MFCC, PLP, PNCC (*power normalized cepstral coefficients*) [23], tandem feature [24, 25], and the recently proposed ICMC (*infinite impulse response constant Q mel-scaled cepstral coefficients*) [26] were used at the front-end. Among these, MFCC remains the most commonly used acoustic features. Also, a vast majority of our sub-systems use energy-based voice activity detector (VAD) in view of its simplicity and effectiveness. Other options for VAD that have been adopted are (i) VQ-VAD [21] in Sys1 and Sys14, (ii) speech/non-speech probabilities inferred from the DNN senone posterior in Sys9, and (iii) two-channel VAD [22] in Sys12.

***Increased Complexity***: We observed a general trend in using higher dimensional feature vector and a larger UBM. From Table 3, the input dimensionality of acoustic features ranges from 39 to 139 with majority of the systems settle at 60. The so-called *tandem features* used in Sys9 and Sys11 were formed by concatenating MFCC and *deep bottleneck feature* (DBF) lead-

Table 3: *List of 17 systems used to form the I4U primary submission to SRE'16, their key features at the front-end, the classifier, unlabeled data, and toolkits that have been used.*

| Sys | Features, dim | VAD | UBM type, size | IV dim | Unlabeled data | Classifier | Toolkit |
|---|---|---|---|---|---|---|---|
| 1 | PNCC+$\Delta$ + $\Delta\Delta$, 39 | VQ-VAD [21] | GMM, 512 | 400 | UBM, t-norm<br>Mean replace | PLDA | Alize, Bob |
| 2 | MFCC+$\Delta$ + $\Delta\Delta$, 60 | Energy-based | DNN, 3859 | 600 | SVM -ive samples | SVM | Kaldi, LibSVM |
| 3 | MFCC+$\Delta$ + $\Delta\Delta$, 60 | Energy-based | GMM, 2048 | 600 | Mean subtraction | PLDA | Kaldi |
| 4 | MFCC+$\Delta$ + $\Delta\Delta$, 60 | Energy-based | GMM, 2048 | 600 | Mean subtraction | PLDA | Kaldi, MSR toolkit |
| 5 | MFCC+$\Delta$ + $\Delta\Delta$, 60 | Energy-based | GMM, 2048 | 600 | Mean subtraction, SVDA | PLDA | Kaldi |
| 6 | ICMC+$\Delta$ + $\Delta\Delta$, 60 | Energy-based | GMM, 2048 | 600 | - | PLDA | Inhouse |
| 7 | MFCC+$\Delta$ + $\Delta\Delta$, 39 | Energy-based | GMM, 4096 | 600 | Mean subtraction | 2Cov, PLDA | SideKit, BOSARIS |
| 8 | MFCC+$\Delta$ + $\Delta\Delta$, 60 | Energy-based | GMM, 2048 | 600 | IDVC | PLDA | Kaldi |
| 9 | Tandem, 137 | DNN Posterior | DNN, 2395 | 400 | IDVC | PLDA | Kaldi, SideKit |
| 10 | MFCC+$\Delta$ + $\Delta\Delta$, 60 | Energy-based | DNN, 2304 | 400 | Mean replace | PLDA | SideKit, Theano |
| 11 | Tandem, 90 | Energy-based | GMM, 2048 | 400 | IDVC | PLDA | Kaldi |
| 12 | MFCC+$\Delta$ + $\Delta\Delta$, 60 | Two-channel VAD [22] | GMM, 512 | 300 | UBM, t-norm, z-norm<br>Mean subtraction | PLDA | Inhouse |
| 13 | MFCC+$\Delta$ + $\Delta\Delta$, 60 | Energy-based | GMM, 1024 | 600 | IDVC | PLDA | MSR toolkit |
| 14 | MFCC+$\Delta$ + $\Delta\Delta$, 39 | VQ-VAD [21] | GMM, 2048 | 600 | Mean replace | PLDA | Inhouse |
| 15 | PLP+$\Delta$ + $\Delta\Delta$, 50 | Energy-based | GMM, 512 | 600 | Mean replace | PLDA | Inhouse |
| 16 | MFCC+$\Delta$ + $\Delta\Delta$, 60 | Energy-based | DNN, 2304 | 600 | Mean subtraction | PLDA | SideKit |
| 17 | MFCC+$\Delta$ + $\Delta\Delta$, 60 | Energy-based | GMM, 4096 | 400 | IDVC | PLDA | Alize |

ing to a dimensionality of around a hundred. The size of the UBM increases tremendously from the typical value of 512 in SRE'12 [5] to 2048 being commonly used in SRE'16. Notably, those large UBMs used in Sys2, Sys9, Sys10, and Sys16 are based on DNN and trained in a supervised manner to predict senone targets. Compared to previous evaluations, the increased complexity was made possible by advances in deep learning and also the availability of increasingly powerful computers.

***Embracing Deep Learning***: An emerging trend that we have observed is the adoption of deep learning technique. Though we did not observe a remarkable performance boost in general, as in other speech applications [14, 15, 27], there are a handful of our sub-systems (six out of seventeen in Table 3) that have successfully incorporated deep learning in one form or another: (i) *Deep bottleneck feature (DBF)* in Sys9, (ii) *Stacked bottleneck feature* in Sys11, (iii) *DNN posterior* in Sys2, Sys9, Sys10, Sys16, (iv) *Splice time delay DNN (TDNN)* [16] in Sys2, and (v) *Denoising autoencoder* in Sys14. For the bottleneck features in Sys9 we used a DNN with seven hidden layers each having 1024 hidden units except for the third layer with only 80 units. The DBF was extracted from this layer and concatenated with MFCC forming the tandem feature. The stacked bottleneck features used in Sys11 were obtained by feeding the DBF to another DNN [28, 29]. Another use of DNN is to replace the GMM posterior for frame alignment as proposed in [7, 8, 9]. This was adopted in four sub-systems as shown in Table 3, where the entries to the UBM type are denoted as DNN. Finally, Sys14 explores the idea of transforming i-vectors of noisy short utterances to clean long utterances using denoising autoencoder [31]. The implementation details can be found in I4U system description [32] and the references therein.

***Dataset Shift***: The DNNs used to generate DBF and frame posterior probabilities were trained in supervised manner, which required orthographic transcription. As shown in Table 2, transcribed data is only available in English which exhibits a slightly difference phonetic coverage than the development and test sets in Cebuano/Tagalog and Mandarin/Cantonese. Also, the PLDA trained on the *Switchboard* and *Mixer* datasets might have modeled a distribution different from that of the test i-vectors. This could be cast as a dataset shift problem. One simple solution is to replace the mean vector of the PLDA with the global mean estimated from the unlabeled in-domain data (see Table 2), or equivalently, subtract the global mean (estimated from the unlabeled set) from the test i-vectors. A more elaborate solution is the IDVC (*inter-dataset variability compensation*) [4] which removes the low-rank subspace corresponds to the local means estimated from various datasets. IDVC was implemented in Sys8, Sys9, Sys11, Sys13, and Sys17. Another effective way of using the unlabeled data is score normalization as in Sys1 and Sys12. In addition, *support vector discriminant analysis* (SVDA) [18] performed well in Sys5 for compensating this mismatch by incorporating unlabeled in-domain data without using any pseudo labels.

***Tools***: Also listed in Table 3 are various open-source toolkits used in developing the I4U systems. More than half of the systems use the Kaldi toolkit. Apart from i-vector extraction, Kaldi was also used to implement and train DNN for extracting bottleneck features and senone posteriors. Readily available Kaldi scripts tailored for SRE tasks have made fast prototyping relatively easier than before and therefore promote wider adoption. Other popular tools utilized by I4U consortium members are *SideKit* [33], *Alize* [34], *MSR toolkit* [35], and *BOSARIS* [36] which were developed more specifically for speaker recognition, and *LibSVM* [37], *Bob* [38], Theano [39] designed for general machine learning.

## 4. System and fusion performance

Considering that the I4U consortium had in total 32 base classifiers, an idea to use a classifier selection was clear from the beginning. To that end, we performed initially a 5-fold cross-validation (CV) on the development set. Base classifier selection and whether pre-calibration of the scores was to be performed were decided using a 5-fold CV setting (recording average performance over folds). Pre-calibration showed systematically better results than without pre-calibration. In the selection setting, we experimented with many different ideas on how to fix the final subset of classifiers (noting that we started with 32 base classifiers). Final ensemble was selected using heuristic rule of using the best single systems ($\min C_{\text{primary}} < 0.7$) and one system from each site. This resulted in better cross-validated results than using just the best systems. The base classifier results on Eval set are shown in Table 4. We show separately the performance on Tagalog (tgl) and Cantonese (yue),

Table 4: *Base classifier performance and their fusions in terms of equal-error-rate (EER%) on Eval and Dev sets, shown separately for Tagalog (tgl), Cantonese (yue), Cebuano (ceb) and Mandarin (cmn) trials.*

| | Eval set | | Dev set | |
|---|---|---|---|---|
| Sys. | EER (tgl) | EER (yue) | EER (ceb) | EER (cmn) |
| 1 | 36.96 | 36.85 | 24.00 | 13.82 |
| 2 | 16.96 | 9.13 | 23.33 | 12.34 |
| 3 | 18.16 | 9.46 | 22.33 | 12.84 |
| 4 | 17.28 | 8.46 | 21.78 | 9.75 |
| 5 | 15.85 | 7.10 | 22.50 | 9.83 |
| 6 | 21.13 | 11.05 | 26.80 | 14.47 |
| 7 | 17.38 | 8.81 | 23.35 | 10.68 |
| 8 | 16.67 | 7.83 | 20.66 | 9.50 |
| 9 | 16.21 | 7.62 | 24.08 | 10.79 |
| 10 | 20.68 | 11.98 | 27.63 | 13.79 |
| 11 | 18.30 | 9.07 | 25.00 | 10.12 |
| 12 | 19.78 | 13.23 | 24.83 | 13.13 |
| 13 | 21.20 | 12.71 | 27.05 | 14.47 |
| 14 | 36.80 | 32.13 | 23.66 | 13.73 |
| 15 | 19.83 | 9.70 | 22.25 | 12.38 |
| 16 | 19.78 | 11.06 | 27.82 | 13.33 |
| 17 | 16.73 | 7.91 | 23.04 | 10.39 |
| Primary | **12.94** | **5.03** | 17.70 | 6.70 |
| Cont. 1 | 13.54 | 6.22 | 17.19 | 6.83 |
| Cont. 2 | 13.93 | 5.60 | **15.53** | **5.90** |

and notice immediately that Tagalog is remarkably harder set than Cantonese. Majority of the base classifier obtain around 20% EER on Tagalog and around 10% EER on Cantonese. Minimum is 15.85% and 7.10% EER for Tagalog and Cantonese, respectively. Also shown in the table are the EERs on Dev set by which the hyper-parameters for individual sub-systems were optimized. The Eval set is generally easier than the Dev set, where the EERs on Tagalog (tgl) and Cantonese (yue) are lower compared to Cebuano (ceb) and Mandarin (cmn), respectively. This is true for all base classifiers except for Sys1 and Sys14, which appear to be over-fit on the Dev set. Despite some base classifiers performing two times worse than the single best classifier, the fusion results consistently outperform the single best, which signifies the benefit of *megafusion* of a large ensemble of classifiers.

***The primary submission*** fusion was designed so that we first pre-calibrated all 17 base classifiers, via minimizing the $C_{\text{wlr}}$ cost, with $p_{\text{tar}} = 0.01$. Development set was used to estimate the scale and bias for all the base classifiers and then applied to the eval-set scores. Then the linear fusion parameters were estimated on the pre-calibrated scores using the same settings. We notice in the Table 4 that this fusion strategy significantly improves in terms of EER compared to any single best base classifier.

***The first contrastive*** system was based on the approach to focus on a smooth transition from minor to major languages. It consists of a trial-wise unsupervised calibration followed by a simple score averaging. To this end, we divided the unlabeled data (only the major language partition was used) into 5 subsets: 4 high-confidence language-gender subsets that account for about 60% of the data and a single subset containing the remaining utterances to be discarded. The purpose of our language-gender subsets was two-fold. Firstly, we used pairwise scores within each subset to train Gaussian classifiers for target and non-target distributions. Secondly, we used it to find subset-dependent log-likelihoods for each enrolment and test i-vectors separately. Summed log-likelihood scores for all i-vectors within each trial was used to determine the subset label of a trial and apply the corresponding calibration parameters.

Table 5: *Performance of the primary and contrastive fusions on Eval set in terms of EER, Minimum and Actual $C_{primary}$. Each entry represents the **Equalized** and **Unequalized** performance metrics [3].*

| | Evalset | | |
|---|---|---|---|
| **Equalized** | EER (%) | $C_{primary}$ (Min) | $C_{primary}$ (Act) |
| Primary | 8.59 | 0.6392 | 0.8779 |
| Cont. 1 | 9.58 | **0.7538** | **0.7615** |
| Cont. 2 | 9.28 | 0.7118 | 2.5538 |
| **Unequalized** | EER (%) | $C_{primary}$ (Min) | $C_{primary}$ (Act) |
| Primary | 8.80 | 0.6328 | 1.0617 |
| Cont. 1 | 10.00 | **0.7473** | **0.8410** |
| Cont. 2 | 9.60 | 0.7218 | 3.2037 |

***The second contrastive*** system was based on the fusion of 32 base classifiers, where unwanted base classifiers were automatically removed. The scores were subjected to z-normalization, where the z-norm parameters were learned from the dev set. We applied the OSCAR [40] sparsity promoting regularization to the $C_{\text{wlr}}$ objective. We notice in Table 4 that OSCAR based fusion system reaches the lowest EER on dev set, but on eval set the primary submission has a better performance.

***Calibration performance*** of the primary and contrastive fusions are shown in Table 5. We contrast their performance in terms of *minimum* and *actual* DCF on the Eval set. We notice that only in Contrastive 1 fusion we achieved low calibration error (i.e., actual DCF is closer to the minimum DCF values) on Eval set which indicates a more effective use of unlabeled data in score calibration. We note that most of the base classifiers did not use unlabeled set in score normalization.

## 5. Conclusions

This paper presents an overview of the recognition systems and their fusion developed for NIST SRE'16 by I4U consortium. The collaboration of 62 researchers from 16 research teams benefited all members during the preparation of robust speaker recognition systems. I4U submissions to SRE'16 encompass 32 sub-systems, each one of them presenting a high-end system involving several weeks or months of careful parameter optimization and data engineering. Even if such massive megafusion may be challenging to apply in real use cases, shared evaluation resources and fusion methodology are important tools to enable large-scale collaboration; it facilitates attacking a common challenge engineering goal with shared resources.

NIST SRE'16 was positioned to tackle a more open-ended problem commonly encountered in practical deployment – language mismatch, lack of labeled data, fast adaptation using limited/unlabeled data. To this end, dataset shift could be accounted for with mean compensation. It is also clear that unlabeled dataset could be used for UBM training and score normalization, which does not required labels.

## 6. Acknowledgements

This work was partially funded by the German BMBF within the CRISP center and the Hesse state (project no. 467-15/09, 518/16-30), and by the European Commission in its framework programme Horizon 2020 under grant agreement no. 706668 (Talking Heads) and OCTAVE Project (#647850). The views expressed in this paper are those of the authors and do not engage any official position of the funding agencies.

# 7. References


[1] A. F. Martin and J. S. Garofolo, "NIST speech processing evaluations: LVCSR, speaker recognition, language recognition," in *Proc. IEEE Workshop on Signal Processing Applications for Public Security and Forensics*, 2007, pp. 1 – 7.

[2] "The NIST Year 2012 Speaker Reecogition Evaluation." [Online]. Available: https://www.nist.gov/multimodal-information-group/speaker-recognition-evaluation-2012

[3] "NIST 2016 Speaker Reecogition Evaluation." [Online]. Available: https://www.nist.gov/multimodal-information-group/speaker-recognition-evaluation-2012

[4] H. Aronowitz, "Compensating inter-dataset variability in PLDA hyper-parameters for robust speaker recognition," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2014, pp. 282–286.

[5] R. Saeidi, K. A. Lee, T. Kinnunen *et al.*, "I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification," in *Proc. INTERSPEECH*, Aug. 2013, pp. 1986 – 1990.

[6] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, pp. 19–41, 2000.

[7] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. IEEE ICASSP*, 2014, pp. 1695–1699.

[8] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition," in *Proc Odyssey: The Speaker and Language Recognition Workshop*, 2014, pp. 293–298.

[9] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[10] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[11] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE Computer Vision*, 2007, pp. 1–8.

[12] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2010, p. 14.

[13] K. A. Lee, A. Larcher, C. You, B. Ma, and H. Li, "Multi-session PLDA scoring of i-vector for partially open-set speaker detection," in *Proc. Interspeech*, 2013, pp. 3651–3655.

[14] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30 – 42, Jan. 2012.

[15] G. Hinton, L. Deng, D. Yu, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. S. abd George Dahl, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82 – 97, Nov. 2012.

[16] D. Snyder, D. Garcia-Romero, and D. Povey, "Time delay deep neural network-based universal background models for speaker recognition," in *2015 IEEE ASRU*. IEEE, 2015, pp. 92–97.

[17] D. Yu and M. L. Seltzer, "Improved bottleneck features using pre-trained deep neural networks," in *Proc. INTERSPEECH*, 2011, pp. 237 – 240.

[18] F. Bahmaninezhad and J. H. Hanesn, "i-vector/PLDA speaker recognition using support vectors with discriminant analysis," in *IEEE ICASSP*, 2017.

[19] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *Proc. IEEE ICASSP*, 2013, pp. 7649–7653.

[20] P. Kenny, T. Stafylakis, P. Ouellet, and M. J. Alam, "JFA-based front ends for speaker recognition," in *Proc. IEEE ICASSP*, 2014.

[21] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data." in *Proc. IEEE ICASSP*, 2013, pp. 7229 – 7233.

[22] M. W. Mak and H. B. Yu, "A study of voice activity detection techniques for NIST speaker recognition evaluations," *Computer, Speech and Language*, vol. 28, no. 1, pp. 295–313, Jan 2014.

[23] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4101–4104.

[24] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. IEEE ICASSP*, 2000, pp. 1635 – 1639.

[25] F. Grézl, M. Karafiát, S. Kontár, and J. Černocký, "Probabilistic and bottleneck features for LVCSR of meetings," in *Proc. IEEE ICASSP*, 2007, pp. 757 – 761.

[26] H. Delgado, M. Todisco, M. Sahidullah, A. Sarkar, N. Evans, T. Kinnunen, and Z.-H. Tan, "Further optimisations of constant q cepstral processing for integrated utterance verification and text-dependent speaker verification," in *Proc. Odyssey*, 2016.

[27] K. A. Lee, H. Li, L. Deng, V. Hautamaki *et al.*, "The 2015 NIST Language Recognition Evaluation: the Shared View of I2R, Fantastic4 and SingaMS," in *Proc. Interspeech*, 2016, pp. 3211–3215.

[28] F. Grezl, M. Karafiat, and K. Vesely, "Adaptation of multilingual stacked bottle-neck neural network structure for new language," in *Proc. IEEE ICASSP 2014*, 2014, pp. 7654–7658.

[29] H. H. Xu, V. H. Do, X. Xiao, and E. S. Chng, "A comparative study of BNF and DNN multilingual training on cross-lingual low-resource speech recognition," in *Proc. of Interspeech 2015*, Dresden, Germany, Sep. 2015.

[30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget *et al.*, "The Kaldi speech recognition toolkit," in *Proc IEEE ASRU*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[31] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1096–1103.

[32] K. A. Lee *et al.*, "The I4U submission to the 2016 NIST speaker recognition evaluation," in *Proc. NIST SRE 2016 Workshop*, 2016.

[33] A. Larcher, K. A. Lee, and S. Meignier, "An extensible speaker identification sidekit in python," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5095–5099.

[34] A. Larcher, J.-F. Bonastre, B. Fauve, K. A. Lee, C. Lévy, H. Li, J. S. Mason, and J.-Y. Parfait, "ALIZE 3.0 - Open Source Toolkit for State-of-the-Art Speaker Recognition," 2013, pp. 2768–2773.

[35] S. O. Sadjadi, M. Slaney, and L. Heck, "MSR identity toolbox v1. 0: A matlab toolbox for speaker-recognition research," *Speech and Language Processing Technical Committee Newsletter*, vol. 1, no. 4, 2013.

[36] N. Bümmer and E. de Villiers, "The BOSARIS Toolkit User Guide: Theory, Algorithms and Code for Binary Classifier Score Processing," AGNITIO Research, South Africa, Tech. Rep., 2011.

[37] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[38] A. Anjos, L. El-Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel, "Bob: a free signal processing and machine learning toolbox for researchers," in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 1449–1452.

[39] The Theano Development Team, "Theano: A python framework for fast computation of mathematical expressions," *arXiv preprint arXiv:1605.02688*, 2016.

[40] S. Petry and G. Tutz, "The OSCAR for generalized linear models," *Technical Report*, 2011.