UTD-CRSS SYSTEMS FOR 2018 NIST SPEAKER RECOGNITION EVALUATION

Chunlei Zhang[§]*, Fahimeh Bahmaninezhad*[§]*, Shivesh Ranjan*[§]*,*
Harishchandra Dubey, Wei Xia, John H.L. Hansen[♠]

Center for Robust Speech Systems (CRSS), University of Texas at Dallas, Richardson, Texas, USA
{chunlei.zhang,fahimeh.bahmaninezhad,shivesh.ranjan,john.hansen}@utdallas.edu

ABSTRACT

In this study, we present systems submitted by the Center for Robust Speech Systems (CRSS) from UTDallas to NIST SRE 2018 (SRE18). Three alternative front-end speaker embedding frameworks are investigated, that includes: (i) i-vector, (ii) x-vector, (iii) and a modified triplet speaker embedding system (t-vector). Similar to the previous SRE, language mismatch between training and enrollment/test data, the so-called domain mismatch, remains as a major challenge in this evaluation. In addition, SRE18 also introduces a small portion of audio from an unstructured video corpus in which speaker detection/diarization is supposedly needed to be effectively integrated into speaker recognition for system robustness. In our system development, we focused on: (i) building novel deep neural network based speaker discriminative embedding systems as utterance level feature representations, (ii) exploring alternative dimension reduction methods, back-end classifiers, score normalization techniques which can incorporate unlabeled in-domain data for domain adaptation, (iii) finding an improved data set configurations for the speaker embedding network, LDA/PLDA, and score calibration training (v) and finally, investigating effective score calibration and fusion strategies. The final resulting systems are shown to be both complementary and effective in achieving overall improved speaker recognition performance.

Index Terms— speaker recognition, speaker embedding, deep neural network, NIST SRE, domain adaptation

1. INTRODUCTION

The focus of 2018 version of NIST SRE (SRE18) is speaker detection, which is consistent with the previous SREs, i.e., to determine whether a specified target speaker is speaking during a given segment of speech or not. SRE18 is organized in a similar manner to SRE16, focusing on speaker recognition over conversational telephone speech collected outside North America. To this end, Call My Net 2 (CMN2) and Video Annotation for Speech Technology (VAST) corpora are utilized as the SRE18 development (DEV) and evaluation (EVAL) sets to support speaker recognition research [1]. As a continuation of SRE16 [2], the CMN2 part of SRE18 follows most of SRE16's evaluation setup, e.g., (1) a huge amount of English training data, while a small volume of in-domain non-English DEV set; (2) duration varies from 10 to 60 seconds for the test cuts. In addition to these similarities, there is a major difference between SRE18 and SRE16. In SRE18, the DEV trials are generated from the same Tunisian Arabic language as for final EVAL trials. This change makes system development experiments being able to mirror the EVAL conditions, whereas SRE16 DEV trails and EVAL trials are language independent [2, 1, 3, 4].

For the VAST part, there are several significant differences: (1) each recording may contain speech from multiple talkers, and manually produced diarization labels are provided for both DEV and EVAL enrollment utterances, but not for test cuts; (2) duration of the samples vary from a few seconds to several minutes; (3) sample rate is higher than CMN2 (i.e., 44 kHz, in stead of 8 kHz) (4) VAST DEV trial list is extremely small (270 trials), which creates challenges in almost all the stages of system development.

In this paper, we present the UTDallas-CRSS solutions to SRE18. Given the two very different evaluation conditions of CMN2 and VAST, we take different approaches for system development. In Sec.2, we describe several baseline speaker embedding systems where different dataset configurations, acoustic features, frameworks for training speaker embeddings are examined towards complementary individual systems; Sec.3 introduces the major back-end solutions for SRE18, including the unsupervised/supervised domain adaptation at the dimension reduction or scoring level, as well as the observations toward better score calibration and fusion strategies. Sec.4 details each of the UTDallas-CRSS sub-systems, the score calibration and fusion methods for CMN2 and VAST, and the formation of CRSS final evaluation submissions to NIST SRE18; Sec.5 reports CRSS primary system performance on SRE18 DEV set and EVAL set, the result of SRE16 EVAL set is also included as reference. Finally, we conclude our work with future direction towards SRE18 in Sec.6.

2. CRSS SPEAKER EMBEDDING SYSTEMS

Three general speaker recognition frameworks were explored for SRE18 submission. The first one is a traditional UBM/i-vector system which was effective in NIST SRE 2016 [3, 5], the second one is a x-vector speaker embedding system adopted from Kaldi [6, 7, 8], and the third one is an improved triplet loss (named t-vector, where an additional L_2 constrained softmax loss term is introduced to formulate a multi-task learning objective) based speaker embedding system [9, 10, 11, 12], see Sec.2.3 for more details.

For submissions to NIST, only the fixed condition is considered at CRSS systems. Even just for fixed condition, there are still a huge amount of training data available for system development. To better clarify the training data usage, Table 1 summarizes the different configurations w.r.t. corresponding frameworks.

Here, SWB covers all Switchboard II phase 2 & 3 and Switchboard Cellular Part 1 & 2 corpora. Three datasets listed in Table 1

[§]The first three authors have equal contributions.

[♠]This project was funded in part by AFRL under contract FA8750-15-1-0205 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen.

Table 1. Corpora used in the speaker embedding system training.

dataset	copora	min-utt/spk	#spk	system
D1	SRE04-08, SWB	1	5756	t-vec,i-vec
D2	D1+Mixer 6	8	4867	t-vec,x-vec
D3	D2+SRE10, voxceleb1	8	6197	x-vec
D4	D3	4	6812	x-vec
D5	voxceleb1 & 2	4	7244	t-vec,i-vec,x-vec

(i.e., D2, D3, D4) are augmented by 3-folds after convolving with far-field Room Impulse Responses (RIRs), or by adding noise from the MUSAN corpus [13]. Kaldi x-vector recipe is adopted for this portion of process. Different speaker filtering criterion is applied to different training datasets. For example, 8 min-utt/spk stands for the filtering process that all speakers with less than 8 utterances and less than 500 frames per utterance were excluded for training. After the speaker filtering, the number of speakers for each dataset and what systems use them for training are listed in Table 1. Three different speaker embedding frameworks are described in the following sections.

2.1. CRSS1: UBM/i-vector system

The i-vector models achieved great success in the past SREs [5, 14, 3, 4]. Among different methodologies, the UBM based i-vector framework was approved to be the most effective in the SRE16 with language domain mismatch [5, 14, 3, 4]. For this consideration, an UBM/i-vector system is developed for SRE18. In this framework, we extract 60 dimensional features (20-D MFCC and $\Delta + \Delta\Delta$) on a 25ms window, with a shift size of 10ms. Non-speech frames are discarded using an energy-based voice activity detection (VAD). In addition, cepstral mean normalization is applied with a 3-second sliding window. 2048-mixture full covariance UBM and total variability matrix is trained using the data listed in Table 1. It is noted that for VAST part alone, we train a second i-vector system using Voxceleb 1 & 2 (D5, down-sampled to 8 kHz) but keep the model parameters the same with the D1 version, for the hope that Youtube microphone data could provide complementary information to the SRE+SWB model.

2.2. CRSS2: x-vector system

The x-vector has been reported to achieve the state-of-the-art speaker recognition performance in recent studies [7, 8]. The model is a deep neural network (DNN) based speaker discriminative framework with practical techniques such as speech segmentation, data augmentation and statistical pooling etc. The network is trained with a softmax loss function and corresponding speaker labels, given by Equation 1:

$$\mathcal{L}_s = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T f(\mathbf{x}_i)+b_{y_i}}}{\sum_{j=1}^{C} e^{W_j^T f(\mathbf{x}_i)+b_j}}, \tag{1}$$

where N is the batch size, C is the total speaker number in the training set, $f(\mathbf{x}_i)$ is the output of the embedding layer of the network (i.e., speaker embedding). y_i is the corresponding class label, and W and b are the weights and bias for the last softmax layer of the network which acts as a classifier.

To process the CMN2 and VAST trials, different x-vector based systems were experimented with a way of incrementally adding the data sets. We used the standard Kaldi x-vector recipe to train the different systems. The dataset configurations for training our 3 x-vector extractors as described in Table 1.

We also trained a separate x-vector system to handle the VAST trials exclusively. To this end, we used only Voxceleb1 & 2 data at 16 kHz to formulate the x-vector system [15, 16].

2.3. CRSS3: t-vector system

Triplet loss is another popular objective function for training face or speaker verification systems [17, 9]. The t-vector system that we have developed for SRE18 is modified from [11], with the changes in loss function and acoustic features.

2.3.1. High resolution filter bank features

High resolution filter bank features are adopted for system development. At the frequency axis, 96-D log mel filter bank features are extracted from a 32ms speech frame, with a 50% overlap between neighboring frames. None-speech part of the utterance is removed by an energy based VAD. To deal with the long duration samples in SRE and SWB data, we uniformly segment the speech utterance into 12-second trunks without overlapping, which is equivalent to 750-D in the time axis as the input to the network. To estimate the embedding at the utterance level, we perform segment level embedding average in a sequential order, there we arrive at t-vector.

The same Inception-resnet-v1 network is employed for speaker discriminative training. To validate the training progress, SRE10 10s-10s condition is employed for this purpose. With the modification just in the feature extraction and keeping everything the same with [11] (i.e., same network and triplet loss function), we can observe +0.5% absolute EER improvement on the SRE10 10s-10s trials.

2.3.2. A multi-task training objective

Inspired by the success of the softmax loss in x-vector models, we performed a modification at the loss function level for the triplet loss based system. Specifically, we formulate a multi-task learning framework by adding a L_2 normalized softmax loss ($\mathcal{L}_{s_{L_2}}$), which is a simple upgrade of original softmax loss:

$$\mathcal{L}_{s_{L_2}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T f(\mathbf{x}_i)+b_{y_i}}}{\sum_{j=1}^{C} e^{W_j^T f(\mathbf{x}_i)+b_j}}, \tag{2}$$

$$\text{subject to } \|f(\mathbf{x}_i)\|_2 = \alpha, \quad \forall i = 1, 2, ..., N$$

where a simple L_2 normalization is applied to the embedding layer before softmax layer, α is a constant that constrains the radius of the speaker embedding hypersphere. α is set to 24 empirically in our experiments. By this operation, we are able to match between the training and test process (i.e., a L_2-norm embedding layer for softmax training, and the same layer for embedding extraction). The total loss function is an integration of three components: a triplet loss term $\mathcal{L}_{triplet}$, a L_2-norm softmax loss term $\mathcal{L}_{s_{L_2}}$ and a regularization term $\mathcal{L}_r$ which alleviates the over-fitting issue during training.

$$\mathcal{L}_{total} = \mathcal{L}_{triplet} + \omega_1\mathcal{L}_{s_{L_2}} + \omega_2\mathcal{L}_r \tag{3}$$

practically, we find 0.1 and 2e-5 for $\mathcal{L}_{s_{L_2}}$ and the L_2 regularization $\mathcal{L}_r$ is a good combination for most of the SRE experiments.

2.3.3. Triplet sampling and shuffling

With the update in the loss function, one necessary change is also made in the triplet sampling module. Previously in [11], we chose a subset of speakers in the training pool for triplet formulation in each

epoch. With the additional $\mathcal{L}_{s_{L_2}}$, it is better to see all the speakers in one epoch. In the experiments, we always randomly select segments from all training speakers for the triplet generation and shuffling to make sure all classes can be seen within one epoch.

2.3.4. Validation on SRE10 10s-10s trials

The t-vector system is trained on various combinations of dataset, loss function, embedding size and front-end features, with the hope to find the best configuration for SRE18. Table 2 lists the validation performance on SRE10 10s-10s condition.

Table 2. Validation performance on SRE10 10s-10s condition.

dataset	mel fbank	loss	emb-dim	EER %
D1	40-D +Δ+ΔΔ	triplet	400	11.4 [11]
D1	96-D	triplet	256	10.9
D1	96-D	multi-task ($\mathcal{L}_s$)	256	10.1
D1	96-D	multi-task ($\mathcal{L}_{s_{L_2}}$)	256	9.7
D2	96-D	multi-task ($\mathcal{L}_{s_{L_2}}$)	128	9.1
D2	96-D	multi-task ($\mathcal{L}_{s_{L_2}}$)	256	**8.7**
D2	96-D	multi-task ($\mathcal{L}_{s_{L_2}}$)	384	9.8

From the Table 2, we find a system configuration with the multi-task objective with the L_2-norm softmax loss, 96-D log mel fbank features, 256-D embedding size and D2 dataset achieves the best performance. Here, a simple cosine distance scoring is used for illustration. With a better LDA/PLDA back-end classifier, the t-vector model could reach state-of-the-art 7.6 % EER on SRE10 10s-10s condition [7, 11].

3. UTDALLAS-CRSS BACK-END SCORING STRATEGIES

Section 2 presents the details of how speaker embedding systems are developed for SRE18. Due to the significant difference in training the frameworks, very different properties are observed from different speaker embedding systems. For example, the SVDA/LDA/PLDA back-end (Fig.1a) is effective for i-vector based system. However, the supervised/unsupervised adaptation techniques (Fig.1b, which tend to work better with {t,x}-vectors) can not bring the same improvement over the out-of-domain (OOD) PLDA, comparing with the Fig.1a solution on i-vector model. In order to maximize the benefit from each system, we customize effective back-end strategies for different speaker embedding systems as follows.

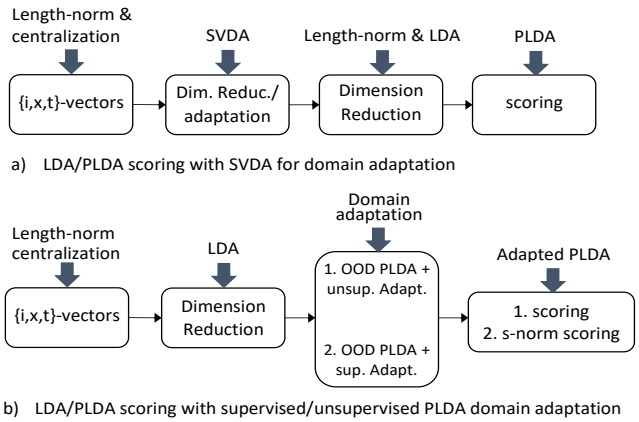

Fig. 1. Flow diagrams of CRSS back-end classifiers

3.1. back-end scoring for i-vector systems

Fig.1a presents the SVDA/LDA/PLDA process which is employed as a back-end classifier for the i-vector model. After extracting i-vectors, the global mean calculated from unlabeled and clean SRE training data is subtracted from all CMN2 and VAST i-vectors, respectively. Next, i-vectors are length-normalized and their dimensionality are reduced from 600 to 350 for CMN2 and 250 for VAST; using SVDA/LDA [18, 19] (i.e., first SVDA reduces the dimension from 600 to 450 for CMN2 and 400 for VAST and then LDA is used to reduce the dimension to 350 and 250 respectively). Length normalization is again applied after dimension reduction. Finally, trial-dependent mean subtraction is employed (i.e., the enrollment and test i-vectors within a trial are averaged and the value is subtracted from the trial-dependent i-vectors) and scores are calculated using PLDA. To train the PLDA model for scoring, the SRE04-08 and SRE18 unlabeled development i-vectors are utilized for CMN2 part, and the phone number is taken as the speaker label for the unlabeled SRE18 DEV data. While for the VAST part, only the SRE04-08 i-vectors are trained to get the PLDA model. The MSR-Identity toolkit is adopted for the back-end implementation [20].

3.2. back-end scoring for x-vector and t-vector systems

For x-vector and t-vector systems which are both discriminatively trained, similar behavior is observed at the scoring. For each embedding system, two alternative scoring methods are proved to be more beneficial than the Fig.1a solution. The same centralization/mean subtraction as the i-vector model is applied to {t,x}-vector models. For the first scoring pipe, LDA is applied to reduce the dimensionality of the embeddings, PLDA with unsupervised parameter adaptation (i.e., mean and variance adaptation using the unlabeled SRE18 DEV data) is followed to get the final scores [21]. For the second scoring pipe, LDA is again applied before PLDA, but in this method, domain adaptation is implemented with the interpolation of OOD PLDA and in-domain PLDA [22], where in-domain PLDA is training with unlabeled SRE18 DEV data and the test date of SITW core-multi condition for CMN2 and VAST [23], respectively. Speaker clustering is performed to get the labels for in-domain PLDA training. Also, score normalization (s-norm) is used when generating the PLDA score, with adaptive cohort selection scheme followed by top score selection [24]. In particular, cohorts were selected from SRE18 unlabeled DEV set for CMN2 partition. For VAST partition, cohorts were selected from the test data of SITW core-multi condition.

4. UTDALLAS-CRSS SUBMISSIONS

In this section, we list the single system configuration and its performance on SRE18 DEV and EVAL set. Calibration and fusion strategies are thereby evaluated based on the DEV results, which formulate the submissions to SRE18.

4.1. Single system performance

The best system performances from each framework are presented in Table 3. It is noted that most of the complementary information are provided by these three best single systems on SRE18 DEV trials (systems built only with Voxceleb 1 & 2 provide marginal/negative improvement in fusion). As shown from Table 3, x-vector systems produce the best performance, followed by t-vector and i-vector systems. S-norm is able to improve the act-Cprimary without additional calibration, at the cost of increasing the EER and min-Cprimary. This is a useful hint for the system calibration and fusion.

Table 3. The single best system performance (before calibration) w.r.t. different training frameworks, training datasets and back-end classifiers on SRE18 DEV and EVAL set. The result is separated by "/" with "DEV/EVAL" order.

system	configuration		CMN2			VAST		
	dataset	classifier	EER (%)	min-Cprimary	act-Cprimary	EER (%)	min-Cprimary	act-Cprimary
i-vector	D1	Fig.1a	10.91/12.41	0.688/0.757	1.0/1.0	11.11/18.73	0.481/0.698	1.0/1.0
x-vector	D4	Fig.1b	**7.2/8.63**	**0.46/0.549**	43.566/47.911	7.41/15.56	0.56/0.624	14.309/13.675
x-vector(s-norm)	D4	Fig.1b	7.36/8.75	0.529/0.57	0.685/0.833	**8.23/14.44**	**0.407/0.554**	1.914/1.249
t-vector	D2	Fig.1b	**9.60/9.61**	**0.507/0.672**	28.825/27.998	7.82/17.14	0.634/0.718	12.588/13.199
t-vector(s-norm)	D2	Fig.1b	10.12/10.52	0.588/0.682	0.674/0.927	**9.05/15.64**	**0.564/0.624**	2.226/1.432

Table 4. UTDallas-CRSS submissions to NIST SRE18. The numbers in bold are obtained after SRE18 EVAL keys are released.

	system	CMN2			VAST		
		EER(%)	min-Cprimary	act-Cprimary	EER(%)	min-Cprimary	act-Cprimary
DEV	Primary	5.63	0.368	0.388	7.41	0.259	0.296
	Contrastive	5.72	0.372	0.384	7.41	0.296	0.374
EVAL	Primary	7.14(**7.63**)	0.489(**0.490**)	0.537(**0.496**)	16.83(**15.56**)	0.600(**0.592**)	0.608(**0.602**)
	Contrastive	7.14	0.490	0.534	16.51	0.626	0.682

4.2. System calibration

NIST evaluates the team performance based on the act-Cprimary. To this end, score calibration is essential. In our primary submission to NIST, only CMN2 part is calibrated with PAV from BOSARIS toolkit [25], because enough SRE18 CMN2 DEV trails are provided by NIST as the mirror to final EVAL CMN2 set. For the VAST submission, the limited 270-trial VAST list is not a good indicator of act-Cprimary for EVAL VAST trials. We decide to only fuse the s-norm scores (drop the non s-norm systems which have better EER and min-Cprimary) of the VAST trials without calibration, with the hope that a linear score fusion (which can be viewed as a linear calibration) will adjust the nominalized score again.

For contrastive submission, calibration for CMN2 is performed by combining DEV CMN2 trials and unlabeled DEV set trials (speaker labels are estimated by speaker clustering, same as the in-domain PLDA model). For VAST calibration, the trails are generated by concatenating DEV VAST and SITW core-multi trails.

4.3. System fusion

In order to predict final scores combining our multiple single systems. We build a fused model by training two logistic regression models for CMN2 and VAST separately. Let $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ be the features by concatenating of each single system, the target variable y is a Bernoulli random variable for which the probability of occurrence is dependent on the prediction given in Equation 4. Regression coefficients $\boldsymbol{\omega}$ are estimated using the maximum likelihood estimation. Scores from each single system are combined with the estimated coefficients to get the fusion score $\hat{y}$.

$$p(y = 1|\mathbf{x}, \boldsymbol{\omega}) = \frac{1}{1 + \exp\left(-\boldsymbol{\omega}^T \mathbf{x}\right)} \tag{4}$$

$$\hat{y} = \boldsymbol{\omega}^T \mathbf{x} \tag{5}$$

with the linear weights learned, the calibrated CMN2 DEV scores (both s-norm and none s-norm) are integrated as the final score to the evaluation. While for VAST fusion, only the 270 DEV VAST trial scores are directly are trained for fusion.

Table 5. The single best system performance w.r.t. different training framework, and the fusion system performance of SRE16 EVAL.

system	SRE16 EVAL		
	EER (%)	min-Cprimary	act-Cprimary
i-vector	10.95	0.697	0.745
x-vector	8.74	0.561	0.603
t-vector	9.54	0.629	0.643
primary	**6.90**	**0.500**	**0.541**

5. PERFORMANCE OF UTDALLAS-CRSS SUBMISSIONS

Table 4 summarizes the systems that we have submitted to SRE18. In our primary submission, we are able to get consistent EVAL performance as DEV set. The EER for the VAST EVAL set is surprisingly high, which might indicate that: a) VAST is more challenging as the multi-talker issue presents; b) some weak single systems have been fused into the final decision, which degrades the final performance (an improved 15.53 % EER is simply obtained by removing the Voxceleb based systems), c) a more effective speaker embedding training framework is required. One positive side of our primary submission is that low act-Cprimary is achieved, because we decide to fuse only s-norm scores directly and drop the plan of calibrating DEV VAST trials, in which we did for our contrastive VAST submission. Overall, our primary submission could get a mean of act-Cprimary at 0.572. In fact, if we only fuse the s-norm systems for CMN2, 0.496 act-Cprimary is achieved with a 7.6% relative improvement over our primary CMN2 submission, which again proves that s-norm score with a linear fusion is more effective towards a lower min-Cprimary and act-Cprimary gap.

To evaluate the CRSS systems on previous SREs, a side experiment is also conducted on SRE16 just to encourage cross sites reference and reflect recent improvements on the sequential SREs. The best single system performance as well as the fusion result is provided in Table 5. With the SRE18 models, a 6.90 % EER and 0.541 act-Cprimary is reported in this study, which brings approximately 30% EER improvement over our submission to NIST SRE16 [3].

6. CONCLUSION AND FUTURE WORK

In this study, we described single and fused systems submitted from CRSS to NIST SRE18 challenge. A detailed description of CRSS models is presented, including speaker embedding training, back-end classifier development and score calibration and fusion strategies selection. We teamed up with sites like JD.com, Oxfordwav research and joined I4U for SRE18, and were able to produce competitive systems to tackle the emerging issues.

As a review of our efforts to SRE18, there are a few things that we need to pay more attention in future work. For example, the speaker diarization system that has been developed for finding the primary speakers in the recordings seems not working on SRE18 DEV VAST set. During the system development, we were not able to see obvious change in terms of the speaker verification performance on DEV VAST trails. Due to limited trial size and time before the challenge deadline, we decided not to integrate speaker diarization system into our submissions, but it is a good future direction to continue after SRE18.

7. REFERENCES

[1] "NIST 2018 speaker recognition evaluation plan," https://www.nist.gov/sites/default/files/documents/2018/08/17/sre18_eval_plan_2018-05-31_v6.pdf, 2018.

[2] "NIST 2016 speaker recognition evaluation plan," https://www.nist.gov/sites/default/files/documents/itl/iad/mig/SRE16_Eval_Plan_V1-0.pdf, 2016.

[3] C. Zhang, F. Bahmaninezhad, S. Ranjan, C. Yu, N. Shokouhi, and J. H. L. Hansen, "UTD-CRSS systems for 2016 NIST speaker recognition evaluation," in *ISCA INTERSPEECH17*, 2017.

[4] Kong-Aik Lee, Ville Hautamäki, Tomi Kinnunen, Anthony Larcher, C Zhang, A Nautsch, T Stafylakis, G Liu, M Rouvier, W Rao, et al., "The I4U mega fusion and collaboration for NIST speaker recognition evaluation 2016," in *Proc. Interspeech*, 2017.

[5] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech, and Langu. Proce.*, vol. 19, no. 4, pp. 788–798, 2011.

[6] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE, 2011.

[7] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proc. Interspeech*, 2017, pp. 999–1003.

[8] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *IEEE ICASSP*, 2018.

[9] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances," in *ISCA INTERSPEECH*, 2017.

[10] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with flexibility in utterance duration," in *IEEE ASRU*, 2017.

[11] C. Zhang, K. Koishida, and J. H.L. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 9, pp. 1633–1644, 2018.

[12] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[13] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[14] Seyed Omid Sadjadi, Timothée Kheyrkhah, Audrey Tong, Craig Greenberg, Elliot Singer Reynolds, Lisa Mason, and Jaime Hernandez-Cordero, "The 2016 nist speaker recognition evaluation," in *Proc. Interspeech*, 2017, pp. 1353–1357.

[15] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.

[16] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[17] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE CVPR*, 2015.

[18] D. Garcia-Romero and C. Y Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. ISCA INTERSPEECH*, 2011, pp. 249–252.

[19] F. Bahmaninezhad and J. H.L. Hansen, "i-vector/plda speaker recognition using support vectors with discriminant analysis," in *IEEE ICASSP*, 2017, pp. 5410–5414.

[20] 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.

[21] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[22] Daniel Garcia-Romero and Alan McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4047–4051.

[23] Mitchell McLaren, Luciana Ferrer, Diego Castan, and Aaron Lawson, "The speakers in the wild (sitw) speaker recognition database.," in *Interspeech*, 2016, pp. 818–822.

[24] Douglas E Sturim and Douglas A Reynolds, "Speaker adaptive cohort selection for tnorm in text-independent speaker verification," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* IEEE, 2005, vol. 1, pp. I–741.

[25] N. Brümmer and E. de Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf," *arXiv preprint arXiv:1304.2865*, 2013.