INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

SdSV Challenge 2020: Large-Scale Evaluation of Short-duration Speaker Verification

Hossein Zeinali [1,4], *Kong Aik Lee* [2], *Jahangir Alam* [3], *Lukas Burget* [4]

[1] Amirkabir University of Technology, Iran
[2] Institute for Infocomm Research, A*STAR, Singapore
[3] Computer Research Institute of Montreal, Canada
[4] Brno University of Technology, Czech Republic

hzeinali@aut.ac.ir, lee_kong_aik@i2r.a-star.edu.sg, jahangir.alam@crim.ca, burget@fit.vutbr.cz

Abstract

Modern approaches to speaker verification represent speech utterances as fixed-length embeddings. With these approaches, we implicitly assume that speaker characteristics are independent of the spoken content. Such an assumption generally holds when sufficiently long utterances are given. In this context, speaker embeddings, like i-vector and x-vector, have shown to be extremely effective. For speech utterances of short duration (in the order of a few seconds), speaker embeddings have shown significant dependency on the phonetic content. In this regard, the *SdSV Challenge 2020* was organized with a broad focus on systematic benchmark and analysis on varying degrees of phonetic variability on short-duration speaker verification (SdSV). In addition to text-dependent and text-independent tasks, the challenge features an unusual and difficult task of cross-lingual speaker verification (English *vs.* Persian). This paper describes the dataset and tasks, the evaluation rules and protocols, the performance metric, baseline systems, and challenge results. We also present insights gained from the evaluation and future research directions.

Index Terms: Speaker Recognition, Benchmark, Short-duration, Evaluation

1. Introduction

Also known as voice authentication, the purpose of speaker verification is to authenticate a claim of identity using the person's voice. This is realized in the form of an automatic process of either accepting or rejecting the claim by using some salient characteristics inherent in a test utterance. A specificity of speaker verification comes from the possibility of remote authentication through various communication channels such as telephone, voice over IP, or radio-frequency. This potential makes the technology attractive (e.g., transaction authentication for mobile banking) though challenging due to the variety of intrinsic and extrinsic factors (e.g., emotion state, phonetic content, channels, and environments) that can affect speaker verification performance.

Among others, two predominant elements of a spoken utterance are its phonetic content and the vocal characteristic of the speaker. From one sentence to another, or even when the same sentence is uttered by the same speaker, the phonetic variation is an undesirable nuisance to a speaker verification system. Not forgetting as well other extrinsic factors, like channel effects [1], that have to be dealt with delicately. Modern approaches to speaker verification represent speech utterances as fixed-length vectors – the so-called speaker embeddings. From i-vector [2] to the recent x-vector [3] embedding, phonetic variability is suppressed with a simple averaging (i.e., temporal pooling), which has shown to be effective for long utterances.

For speech utterances of short duration (in the order of a few seconds), the speaker embeddings show significant dependency on their phonetic content. In this regard, it is common to model both speaker and spoken content jointly [4], for instance, modeling of speaker pronunciation of individual words, syllables, or phones. This marks the major difference between the text-dependent and text-independent speaker verification tasks [5]. The main idea of the former is to directly exploit the voice individuality associated with a specific phonetic context. Though remains as a subject of much debate, we believe the major impediment lies at the short duration which makes it difficult to suppress phonetic contents from speaker cues via temporal pooling – a core mechanism used in state-of-the-art speaker embedding.

Following the *RedDots Challenge*[1], the *SdSV Challenge*[2] aims to acquire a better understanding and explore new research directions on speaker-phonetic variability modeling for speaker verification over short utterances. Different from the RedDots challenge, the SdSV was organized as an online leader-board challenge. It also features an evaluation set with the number of speakers and trials two-order of magnitude larger than the former [6]. In addition to text-dependent and text-independent tasks, the challenge features an unusual and difficult task of cross-lingual speaker verification (English *vs.* Persian). The three different tasks facilitate systematic benchmark and analysis on varying degrees of phonetic variability (text-dependent, text-independent, and cross-lingual) on short-duration speaker recognition.

This paper describes the SdSV Challenge 2020, the dataset and tasks, the evaluation rules and protocols, the performance metric, baseline systems, and challenge results.

2. Task and Dataset

The SdSV Challenge consists of two separate tasks. We describe below the details of each task and the corresponding dataset for training and evaluation.

2.1. Task Description

Task-1 of the SdSV Challenge is defined as the speaker verification in **text-dependent** mode. It is a twofold verification task in which both speaker and phrase are verified – given a

[1] https://sites.google.com/site/thereddotsproject/reddots-challenge
[2] https://sdsvc.github.io

http://dx.doi.org/10.21437/Interspeech.2020-1485

Table 1: *Number of trials in each partition of Task-1.*

Language	Gender	TC	TW	IC
Farsi	Male	$94,856$	$226,357$	$1,370,326$
Farsi	Female	$154,210$	$368,864$	$2,178,576$
English	Male	$79,804$	$430,933$	$925,257$
English	Female	$133,653$	$721,274$	$1,622,590$
Total		$462,523$	$1,747,428$	$6,096,749$

Table 2: *Number of trials in each partition of Task-2.*

Language	Gender	Target	Imposter
Farsi	Male	$474,920$	$3,008,447$
Farsi	Female	$760,565$	$4,691,234$
English	Male	$278,863$	$1,337,038$
English	Female	$470,659$	$2,176,298$
Total		$1,985,007$	$1,121,3017$

test segment of speech and the target speaker's enrollment data, determine whether the test segment and a specific phrase was spoken by the target speaker. To this end, we define each trial to consist of a test segment along with a model identifier which indicates a phrase ID and three enrollment utterances. The enrollment utterances amount to an average of 7.6 seconds for each speaker-passphrase pair, while the average duration of test utterances is 2.6 seconds.

The enrollment and test phrases are drawn from a fixed set of ten sentences consisting of five Persian and five English phrases. Table 1 shows the number of trials for each language, gender, and trial type. There are four trial types, namely, *target-speaker/correct-phrase* (TC), *target-speaker/wrong-phrase* (TW), *imposter/correct-phrase* (IC) and *imposter/wrong-phrase* (IW) [5]. Systems should only accept TC trials and reject the rest. Since it is not difficult to reject IW trials, this type is not used for the challenge.

Task-2 of the challenge is the speaker verification in text-independent mode. Given a test segment of speech and the target speaker enrollment data, the task is to determine whether the test segment was spoken by the target speaker. Each trial comprises a test segment of speech along with a model identifier which indicates one to several enrollment utterances. The net enrollment speech for each model is randomly distributed between 4 to 180 seconds, after applying an energy-based VAD with an average of 49.3 seconds. The same test data as Task-1 is used here. More information about the challenge, tasks, and rules can be found in the challenge evaluation plan [7].

There are two evaluation conditions for this task corresponding to two types of trials. The first is a typical text-independent trial where the enrollment and test utterances are from the same language (Persian). The second is a text-independent cross-language trial where the enrollment utterances are in Persian and test utterances are in English. Because of the addition of within-speaker cross-language variability, this type of trial is expected to be more difficult. Similar to Task-1, there are no cross-gender trials in Task-2. Table 2 shows the number of trials for each language, gender, and trial type. In order to be able to do a better comparison between text-dependent and independent speaker verification, TC-vs-IC trials for Farsi language from Task-1 were added to Task-2 trials which contain $249,066$ and $3,548,902$ target and imposter trials respectively. In the results tables, this subset is indicated by *FA TC-vs-IC*.

2.2. Training and Evaluation Data

The evaluation and in-domain training data for the challenge are selected from DeepMine dataset [8, 9]. For both tasks a fixed training condition is used where the systems should only be trained using a designated set which composed of VoxCeleb 1&2 [10, 11], LibriSpeech [12] and task-specific in-domain training data.

For the Text-Dependent Task-1, the in-domain training data contains utterances from *963 speakers*, some with only Persian phrases. Model enrollment is done in a phrase and language-dependent way using three utterances for each model. For the Text-Independent Task-2, the in-domain training data in this task contains text-independent Persian utterances from *588 speakers*.

3. Baselines and Performance Metrics

Two baseline systems were made available to the participants illustrating the use of the dataset and the expected results on text-dependent Task-1 and text-independent Task-2. Systems were evaluated based on their detection cost.

3.1. Baselines

We provided two baseline systems. The first baseline is a x-vector system [3], which has shown good performance in a short-duration scenario. The x-vector baseline is denoted as **B01**, and was used for both text-dependent and text-dependent tasks as shown in Section 4. The second baseline is an i-vector/HMM system, denoted as **B02**, and was designed specifically for text-dependent task. The method was proposed in [13, 4] and have achieved very good results on both RSR2015 [5] and RedDots [6] text-dependent datasets.

The x-vector baseline follows an extended-TDNN (E-TDNN) topology [14] and was trained using VoxCeleb 1 and 2 dataset. A linear discriminant analysis (LDA) with 150 dimensions is applied to the extracted x-vectors and after that, a probabilistic LDA (PLDA) with both VoxCeleb datasets is trained. Finally, trials are scored using the PLDA without any score normalization.

Different from that of the x-vector system, the i-vector/HMM baseline considers phrase information and therefore more suitable for the text-dependent task. In this method, i-vector is used as a fixed dimension representation for each utterance. In contrast to the conventional i-vector/GMM method, which uses GMM for aligning input frames to Gaussian components, here monophone HMMs are used for the frame alignment. Therefore, monophone HMMs are trained first using the in-domain training data. Phrase specific HMMs are constructed using the monophone HMMs and used to obtain alignments of frames to HMM states and the GMM components within the states. These alignments are used to extract sufficient statistics from each utterance. Finally, the statistics are used to train an i-vector extractor and to extract i-vectors from enrollment and test utterances. Scoring is done using LDA-Cosine and scores are normalized using the t-norm method. A full description of this method can be found in [4].

3.2. Evaluation Metrics

The main performance metric adopted for the challenge is the normalized minimum *Detection Cost Function* (DCF) defined as a weighted sum of the miss and false alarm error probabilities:

$$C_{Det} = C_{Miss} \times P_{Miss\,|\,Target} \times P_{Target} + C_{FalseAlarm} \\ \times P_{FalseAlarm\,|\,NonTarget} \times (1 - P_{Target}),$$

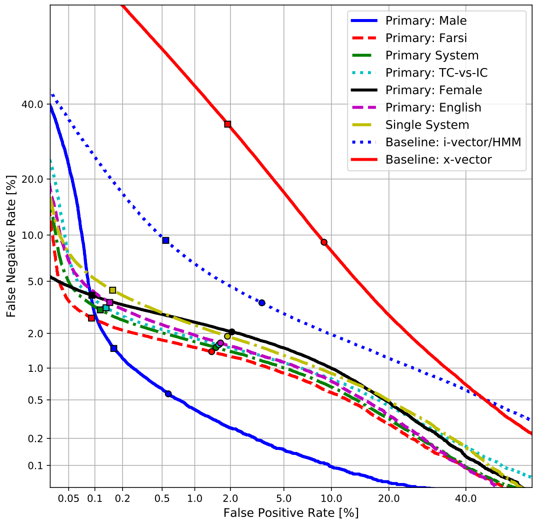

Figure 1: *DET curves of the best performing team for Text-Dependent Task-1.*

where, $C_{Miss} = 10$, $C_{FalseAlarm} = 1$ and $P_{Target} = 0.01$. Based on the parameters, the normalized DCF (DCF_{norm}) is DCF divided by 0.1 as the best cost that could be obtained by constant decision (i.e. rejecting all trials). In addition to MinDCF$^{0.01}_{norm}$, the Equal Error Rate (EER) will be reported as the common metric in speaker verification.

4. Challenge Results

The SdSV challenge was well received by the speaker recognition community. Sixty-seven teams from academia and industry in 23 countries registered for this challenge. Among these, 49 teams registered for both tasks, 4 teams to Task-1 only, and 14 teams enrolled in Task-2 only. At the end of the evaluation phase, we received submissions from 34 teams for Task-2 and 20 teams for Task-1, while there are more submissions on the leader-boards (some teams did not submit the final systems). In this section, we report and discuss the results of this challenge and provide our observations on the results, especially for the best performing teams. Details of the best single and primary systems are reported in their respective papers.

4.1. Progress versus Evaluation Subsets

For the challenge, we created a progress subset and an evaluation subset by taking 30 % and 70 % of the entire trial set, respectively. The progress subset was used for monitoring the performance on the leaderboards, while the evaluation subset was designated for reporting final official results at the end of the evaluation phase.

4.2. Results on Text-Dependent Task-1

Among the 53 registered teams, 20 teams submitted their final scores for Task-1. In Table 3, we report the results of all submitted primary systems of Task-1 in terms of MinDCF and EER. Note that results are reported on both progress and evaluation subsets in *progress/evaluation* format. It could be observed from Table 3 that all participating teams achieved consistent performance across progress and evaluation subsets and no unexpected behavior was observed. This indicates that the subsets of the test trials were created properly. Compared to our i-vector/HMM baseline **B02** 60% of the participating teams provided better performance in both evaluation metrics. All teams except T31 and T65 outperformed the x-vector baseline. The comparison of detection error trade-off (DET) curves between the baselines, primary, and single systems as well as some sub-conditions on the primary system of best performing team (i.e., T56) for Task-1 are illustrated in Figure 1. The definition of the single and primary systems can be found in the challenge evaluation plan [7].

Table 3: *Results of Text-Dependent Task-1. Results are shown in Progress / Evaluation format and sorted based on the MinDCF$^{0.01}_{norm}$ on Evaluation set.*

IDs	EER[%]	MinDCF$^{0.01}_{norm}$	IDs	EER[%]	MinDCF$^{0.01}_{norm}$
T56	1.45/1.52	0.0417/0.0422	T05	2.84/2.94	0.1194/0.1203
T14	1.45/1.52	0.0450/0.0456	B02	3.47/3.49	0.1472/0.1464
T08	1.62/1.69	0.0469/0.0470	T18	4.11/4.21	0.1508/0.1515
T10	1.58/1.60	0.0649/0.0658	T25	6.61/6.61	0.2441/0.2464
T26	2.10/2.14	0.0718/0.0719	T45	5.75/5.77	0.2720/0.2709
T34	2.09/2.13	0.0757/0.0758	T03	6.96/7.01	0.3146/0.3163
T01	2.18/2.23	0.0771/0.0785	T13	13.37/13.49	0.4836/0.4830
T29	2.57/2.61	0.0887/0.0888	T11	8.95/8.98	0.5077/0.5056
T20	2.33/2.34	0.0891/0.0897	B01	9.05/9.05	0.5290/0.5287
T49	3.30/3.37	0.0971/0.0971	T31	9.70/9.63	0.5352/0.5364
T61	2.92/2.96	0.1024/0.1024	T65	9.73/9.67	0.5487/0.5482

Table 4: *Detailed results for the best performing team (i.e. Team56) in Text-Dependent Task-1. The last Best Results columns show the best-achieved performance among all teams for each sub-condition.*

	Single System		**Primary System**		**Best Results**	
Condition	EER	MinDCF	EER	MinDCF	EER	MinDCF
All	1.89	0.0587	**1.52**	**0.0422**	1.52	0.0422
Male	1.02	0.0488	**0.57**	**0.0309**	0.57	0.0309
Female	2.46	0.0647	2.06	**0.0488**	**1.99**	0.0488
Farsi	1.67	0.0471	**1.40**	**0.0357**	1.40	0.0357
English	2.14	0.0721	1.66	**0.0495**	**1.60**	0.0495
TC-vs-IC	2.02	0.0628	**1.61**	**0.0452**	1.61	0.0452
TC-vs-TW	0.05	0.0048	0.06	0.0049	**0.01**	**0.0001**
FA TC-vs-IC	1.73	0.0492	**1.44**	**0.0371**	1.44	0.0371

Table 4 presents the detailed results of the best performing system on the evaluation subset with trials partitioned into different sub-conditions. First of all, female trials appear to be more challenging than male speaker trials. As reported in in [9], speakers with more than one recording device were included in the dataset. The number of female speakers with multiple devices is more than male speakers, which contributes to the huge performance gap. Also, there are more recording sessions for female speakers which means more variations in the female trials. Similar results are observed for text-independent Task-2.

Compared to that of Farsi, English trials appear to be more challenging. The reason is twofold. Firstly, the average duration of English phrases is 20% shorter. Secondly, Farsi is the native language of most speakers in the dataset, and some of them have limited English proficiency.

It is also clear from the results that by having a proper phrase verification system it is not difficult to reject TW trials which is inline with the published results [4, 15]. So, the most important condition in the text-dependent case is rejecting TC trials where impostors try to fool the system by uttering the correct passphrase. One more interesting observation from the results in Table 4 is that the single system gives a competitive performance compared to the primary system. This indicates that a costly and often impractical fusion of multiple subsys-

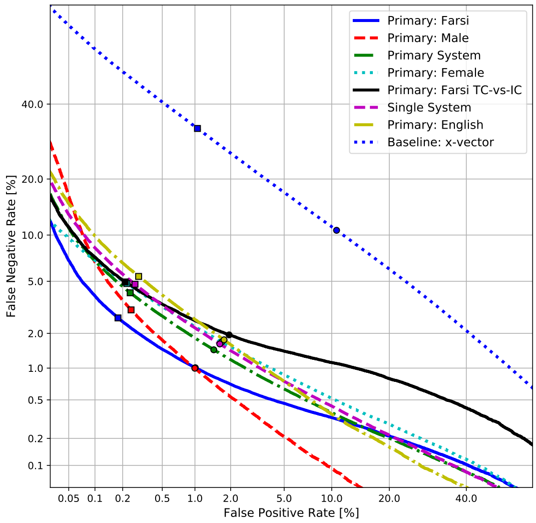

Figure 2: *DET curves of the best performing team for Text-Independent Task-2.*

tems might be unnecessary for practical applications.

4.3. Results on Text-Independent Task-2

In the text-independent Task-2, 63 teams registered for the challenge, and 34 teams submitted their final scores at the end of the evaluation phase. Table 5 shows the results of all primary systems on Task-2 in terms of EER and MinDCF. Except for T63 and T12, all other teams were able to achieve better performance in all evaluation metrics than our x-vector baseline **B01**. Like in Task-1, all systems demonstrated consistent performances across the progress and evaluation subsets. It is worth noticing that the results of top teams show a considerable improvement compared to the baseline.

The DET plots for the baseline system and the primary system as well as the single system of the best performing team (T37) on Task-2 is shown in Figure 2. The DET plots for sub-conditions for the best primary system are also presented in the same plot.

Table 6 reports detailed results of the best-performing team for Task-2 based on different sub-conditions. First of all, similar to Task-1, the performance on male trials is better than female trials due to the mentioned reasons though the difference is not as much compared to Task-1. It seems more enrollment data (possibly from several sessions) reduces the variation effects. The last row of Table 6 shows the results of text-dependent trials which are a subset of Task-1's trials and should be compared with the last row of Table 4. It is obvious that using a specially designed pipeline for the text-dependent task achieves better performance. This has happened while the difficult trial-type TW was eliminated from this comparison because text-independent systems totally failed to reject this kind of trial.

Comparing the results of the typical text-independent speaker verification for Farsi (i.e. fourth row) with cross-language results between Farsi and English shows that speaking in different languages how much affect the performance of the speaker verification. This has happened while the test data for the cross-lingual case is highly accented English (most of the participants speak English like Farsi).

Finally, by comparing the Farsi results of Task-2 (i.e. fourth row) with equivalent results on Task-1 (i.e. fourth row) it is clear that the text-independent performance is better than text-dependent if there is enough enrollment data. Note that this has happened while there are almost seven times more enrollment data in Task-2. So, we can say for limited enrollment data, text-dependent case outperforms text-independent and by increasing the enrollment data we can expect comparable performance for text-independent methods.

Table 5: *Results of Text-Independent Task-2. Results are shown in Progress / Evaluation format and sorted based on the MinDCF$^{0.01}_{norm}$ on Evaluation set.*

IDs	EER[%]	MinDCF$^{0.01}_{norm}$	IDs	EER[%]	MinDCF$^{0.01}_{norm}$
T37	1.45/1.45	0.0654/0.0651	T04	4.46/4.45	0.1942/0.1945
T35	1.51/1.50	0.0745/0.0740	T11	4.46/4.46	0.2020/0.2014
T41	1.77/1.74	0.0770/0.0765	T01	3.96/3.95	0.2078/0.2073
T64	1.84/1.83	0.0839/0.0836	T21	5.55/5.55	0.2352/0.2361
T05	2.00/2.00	0.0957/0.0951	T19	5.69/5.67	0.2369/0.2374
T10	2.32/2.32	0.1048/0.1051	T55	5.72/5.70	0.2530/0.2533
T48	2.68/2.69	0.1122/0.1118	T24	5.98/5.96	0.2638/0.2646
T34	2.74/2.73	0.1171/0.1178	T23	6.61/6.60	0.2679/0.2680
T49	2.85/2.84	0.1252/0.1246	T60	7.13/7.13	0.3074/0.3066
T16	3.07/3.05	0.1263/0.1256	T14	9.04/9.05	0.3378/0.3388
T42	2.89/2.88	0.1264/0.1261	T08	8.48/8.48	0.3398/0.3405
T45	2.95/2.93	0.1262/0.1263	T22	8.20/8.21	0.3548/0.3545
T56	2.71/2.70	0.1317/0.1326	T31	10.04/10.03	0.4194/0.4199
T27	3.03/3.01	0.1377/0.1372	T65	10.60/10.62	0.4287/0.4280
T43	3.67/3.68	0.1576/0.1573	B01	10.67/10.67	0.4319/0.4324
T61	3.95/3.93	0.1643/0.1646	T63	16.24/16.26	0.4785/0.4794
T18	4.47/4.48	0.1885/0.1889	T12	22.16/22.16	0.8929/0.8936
T06	4.66/4.62	0.1918/0.1922			

Table 6: *Detailed results for the best performing team in Task-2 (i.e. Team37). "FA TC-vs-IC" is text-dependent trials from Task-1 for Farsi TC-vs-IC condition.*

	Single System		**Primary System**		**Best Results**	
Condition	EER	MinDCF	EER	MinDCF	EER	MinDCF
All	1.63	0.0742	**1.45**	**0.0651**	1.45	0.0651
Male	1.17	0.0630	1.00	0.0550	**0.98**	**0.0530**
Female	1.88	0.0800	**1.70**	**0.0701**	1.70	0.0701
Farsi	1.12	0.0509	**1.01**	**0.0443**	1.01	0.0443
English Cross	2.03	0.0958	**1.77**	**0.0830**	1.77	0.0830
FA TC-vs-IC	2.19	0.0809	1.94	0.0705	**1.82**	**0.0680**

5. Conclusions

In this work, we presented separately the results on text-dependent and text-independent speaker verification tasks of the recently held *Short-duration Speaker Verification* (SdSV) Challenge 2020. We also summarized and discussed the reported challenge results and provided detailed results of the best performing teams on both tasks. More than 50% of the participating teams' provided systems were able to outperform the baselines for both tasks. One of the captivating observations from this evaluation was that the performance gap between the single and primary fused systems is very narrow, which implies that reliable and competitive text-dependent and text-independent speaker verification systems can be built without applying any fusion strategies.

6. Acknowledgements

The authors would like to thank Sharif DeepMine Ltd. for sponsoring the database, and Computer Research Institute of Montreal (CRIM), and Phonexia Ltd. for sponsoring the prizes.

7. References

[1] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech & Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[2] N. Dehak, P. 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, 2010.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[4] H. Zeinali, H. Sameti, and L. Burget, "HMM-based phrase-independent i-vector extractor for text-dependent speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1421–1435, 2017.

[5] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[6] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. v. Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma *et al.*, "The RedDots data collection for speaker recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "Short-duration Speaker Verification (SdSV) Challenge 2020: the Challenge Evaluation Plan," arXiv preprint arXiv:1912.06311, Tech. Rep., 2020.

[8] H. Zeinali, H. Sameti, and T. Stafylakis, "DeepMine speech processing database: Text-dependent and independent speaker verification and speech recognition in Persian and English." in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 386–392.

[9] H. Zeinali, L. Burget, and J. Cernocky, "A multi purpose and large scale speech corpus in Persian and English for speaker and speech recognition: the DeepMine database," in *Proc. ASRU 2019 The 2019 IEEE Automatic Speech Recognition and Understanding Workshop*, 2019.

[10] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[11] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.

[12] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[13] H. Zeinali, L. Burget, H. Sameti, O. Glembek, and O. Plchot, "Deep neural networks and hidden Markov models in i-vector-based text-dependent speaker verification." in *Odyssey 2016*, 2016, pp. 24–30.

[14] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5796–5800.

[15] H. Zeinali, L. Burget, H. Sameti, and H. Cernocky, "Spoken pass-phrase verification in the i-vector space," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 372–377.