UIAI System for Short-Duration Speaker Verification Challenge 2020

Md Sahidullah, Achintya Kumar Sarkar, Ville Vestman, Xuechen Liu, Romain Serizel, Tomi Kinnunen, Zheng-Hua Tan, Emmanuel Vincent

UIAI SYSTEM FOR SHORT-DURATION SPEAKER VERIFICATION CHALLENGE 2020

Md Sahidullah[1]*, Achintya Kumar Sarkar*[2]*, Ville Vestman*[3]*, Xuechen Liu*[1,3]*, Romain Serizel*[1]*,*
Tomi Kinnunen[3], Zheng-Hua Tan[4], Emmanuel Vincent[1]

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France
[2]Indian Institute of Information Technology, Sri City, India
[3]School of Computing, University of Eastern Finland, Joensuu, Finland
[4]Aalborg University, Aalborg, Denmark

ABSTRACT

In this work, we present the system description of the UIAI entry for the short-duration speaker verification (SdSV) challenge 2020. Our focus is on Task 1 dedicated to text-dependent speaker verification. We investigate different feature extraction and modeling approaches for automatic speaker verification (ASV) and utterance verification (UV). We have also studied different fusion strategies for combining UV and ASV modules. Our primary submission to the challenge is the fusion of seven subsystems which yields a normalized minimum detection cost function (minDCF) of 0.072 and an equal error rate (EER) of 2.14% on the evaluation set. The single system consisting of a pass-phrase identification based model with phone-discriminative bottleneck features gives a normalized minDCF of 0.118 and achieves 19% relative improvement over the state-of-the-art challenge baseline.

Index Terms: Text-dependent speaker verification, Utterance verification, Fusion, Bottleneck feature, SdSV challenge 2020.

1. INTRODUCTION

Automatic speaker verification (ASV) is the task of verifying whether a speech utterance has been spoken by a claimed speaker [1, 2]. State-of-the-art ASV systems show promising performance when several minutes of audio data have been collected in controlled conditions for both enrollment and verification. The amount of enrollment and verification speech is an important factor [3]. While having more speech typically improves recognition performance, short-duration utterances are often preferable for practical deployment. The *short-duration speaker verification* (SdSV) challenge 2020 primarily focuses on the duration factor where speakers are enrolled and verified with a few seconds of audio data [4]. Besides, the speech corpus for the challenge was recorded in realistic environments, and the collection protocol was designed to incorporate various kinds of noises in the speech corpus which introduced mismatches between the enrollment and the verification phases. The challenge has two independent tasks. Our entry focuses on Task 1, which concerns *text-dependent* ASV.

In text-dependent ASV, the spoken phonetic contents for enrollment and verification are assumed to be identical. Typically, a short sentence or phrase is shared by all users. However, considering the practical fact that the test speaker can also utter a wrong phrase [5, 6], there may be four types of trials, respectively defined as *target correct* (TC) where the target speaker utters the correct phrase, *target wrong* (TW) where the target speaker utters a different phrase, *impostor correct* (IC) where an impostor utters the same phrase as in speaker enrollment, and *impostor wrong* (IW) where an impostor utters a different phrase compared to speaker enrollment. For evaluation purposes, TC trials are considered as genuine trials to be *accepted*, and the remaining three as impostor trials to be *rejected*.

Although methods developed for text-independent ASV are applicable to text-dependent ASV without any modification, they do not generalize well [5]. In particular, the performance severely degrades in the TW condition due to the similarities in speaker information. The solutions proposed for text-dependent ASV verify the speaker identity and the spoken phrase in an integrated manner. Here, the phrase information is incorporated by capturing contextual information with a *hidden Markov model* (HMM) [5, 7–10], *dynamic time warping* (DTW) [10, 11] and pass-phrase identification [12]. Alternatively, standalone *utterance verification* (UV) and ASV modules can be developed, and fused together at the score or decision level [13].

We investigate both strategies for this challenge. Our single system applies joint spoken phrase and speaker identity verification where phrase information is incorporated with a *pass phrase-dependent background model* (PBM) [12]. The primary system integrates modules developed for separate tasks. Here, we adopt a *cascade fusion* strategy where UV is performed before ASV. We first compute the decision threshold associated with the *equal error rate* (EER) for UV on the development set. A verification trial is assigned with an

arbitrarily low ASV score if its UV score is lower than the threshold. On the other hand, a trial is passed to the ASV module for scoring if its UV score is greater than or equal to the threshold. The threshold computed on the development set is adopted for scoring on the evaluation set. Our UIAI team is a multi-site collaboration involving four research labs. Given the emphasis on text-dependent ASV, we investigate different UV and text-dependent approaches. We develop an *i-vector*-based UV method that includes channel variability compensation. We introduce a text-dependent ASV method employing *phone-based bottleneck features* with PBM. We also explore how a standard ASV system can be improved in short-duration conditions with different frame-level acoustic features and utterance-level *speaker embeddings*. Finally, we study different system combination strategies suitable for combining UV and ASV systems. Our primary system, which is a multi-level fusion of seven different subsystems, has achieved the fifth rank out of 19 submissions in the challenge whereas the *single system* has shown substantial improvement over the two challenge baselines.

The rest of the paper is organized as follows. Section 2 describes different subsystems developed for the challenge. Section 3 describes the experimental setup. Section 4 presents the experimental results. Finally, we conclude in Section 5.

2. SYSTEM DESCRIPTION

In this section, we summarize the subsystems used for our primary submission to the SdSV challenge.

2.1. Utterance verification

Our UV system relies on speaker embeddings. Although speaker embeddings mainly encode speaker information, they contain a considerable amount of information about the spoken content [14, 15]. This makes them potentially useful for UV tasks (besides ASV). The UV task in the SdSV challenge is a closed-set task as there are no out-of-set phrases. We use an i-vector representation [16] and a *probabilistic linear discriminant analysis* (PLDA) back-end for this task. The setup is similar to the one commonly used in ASV scenarios, except that we treat utterance identifiers (rather than speakers) as the class labels. The i-vectors are projected onto a 9-dimensional space using *linear discriminant analysis* (LDA). Whitening and length normalization are applied to the projected i-vectors. We then use Gaussian PLDA with a full-rank subspace to compute the pairwise UV score between the average i-vector of the claimed phrase and the i-vector of the test phrase. Finally, we apply score normalization suitable for this closed-set scenario where the number of possible hypotheses is fixed. We use Max norm which subtracts the maximum score of the other (competing) phrases from the hypothesized phrase score [13]. We also tried Mean norm but it performed worse.

2.2. Speaker verification

We develop four standalone ASV systems based on *x-vector-PLDA* [17] and *Gaussian mixture model-universal background model* (GMM-UBM) [18] approaches.

X-vector-PLDA system: The network architecture for extracting x-vector speaker embeddings is given in Table 1. It differs from the x-vector architecture originally reported in [17] by adding *squeeze-and-excitation* (SE) [19] modules to each frame-level *time-delay neural network* (TDNN) layer. In addition, three of the frame-level layers are replaced by *residual* (RES) blocks [20]. The global mean and standard deviation pooling layer is replaced with a *learnable dictionary encoder* (LDE) [21, 22]. The LDE layer is similar to a GMM: it assigns features into components and computes statistics locally. We used a variant of LDE with a diagonal covariance matrix shared across all components. For more details of the speaker embedding extractor, see [23]. The trials are scored with a PLDA module [24] trained on the training embeddings. Finally, adaptive symmetric score normalization (AS-norm) [25] is applied.

GMM-UBM systems: The success of GMM-UBM systems in speaker verification with short utterances [3] as well as text-dependent ASV [5, 13, 26] motivates us to explore this approach for the SdSV challenge. Our GMM-UBM systems are similar to standard GMM-UBM systems except that the UBMs are trained in a more efficient way. Instead of training a UBM on the full dataset, we train 10 GMMs on audio data for each of the 10 phrases independently and create the UBM by combining the components of the 10 GMMs and normalizing the mixture weights to unit sum. In the following, we consider three GMM-UBM systems relying on different acoustic features.

Table 1. *Architecture of speaker embedding extractor network. The layers from 1 to 8 and layer 10 are followed by leaky ReLU activations and batch normalization. The embeddings are extracted from layer 10 before applying the activation function.*

#	Layer type	CNN kernel size	Output dim.
1	TDNN-SE	5	512
2,3	TDNN-RES-SE	5	512
4,5	TDNN-RES-SE	7	512
6,7	TDNN-RES-SE	1	512
8	TDNN-SE	1	128
9	LDE aggregation	—	8,192
10	FC	—	512
11	FC-softmax	—	#speakers

2.3. Joint utterance and speaker verification

We perform joint verification of spoken content and speaker identity based on PBMs [12]. In this approach, PBMs are first derived from the GMM-UBM using maximum a posteriori (MAP) adaptation with pass phrase-specific audio data. During the enrollment phase, target speaker models are created by MAP adaptation from the best-matched PBM instead of the single UBM in a GMM-UBM system. The best-matched PBM is found in the maximum likelihood (ML) sense. In the test phase, we first determine the best-matched PBM for the test utterance. Finally, we compute the log-likelihood ratio score between the target model and the best-matched PBM. Our primary submission includes two PBM systems based on two different sets of *bottleneck features* (BN) extracted with deep neural networks (DNNs).

Phone-discriminative bottleneck (Phone-BN) features are extracted by a DNN trained on phone labels obtained using an HMM. We use the in-domain audio data of the SdSV challenge along with phone sequence information. We chose the flat-start method [27] for HMM training as phone-level alignments are not available. First, we train HMM-based *monophone models* by pooling all the utterances and their corresponding phone sequences. Parameters of the phone-based HMMs are then re-estimated in an iterative manner using the *Baum-Welch* algorithm. We consider 42 HMMs that correspond to the number of unique phones in the training data. We consider a 3-state (excluding start and end-state) left-to-right topology without skipping state transitions. We also model silences and pauses between words. We generate phone-level alignments and discard silences and pauses. Then we train a DNN to classify phones. The frame-level output of one hidden layer [28] is then projected using *principal component analysis* (PCA) to obtain lower-dimensional Phone-BN features. Figure 1 illustrates the Phone-BN feature extraction system.

Stream-wise time-contrastive learning based BN (sTCL-BN) features are extracted in a similar way, except that the phone classes used as DNN training targets are found in an unsupervised way [28, 29]. First, all the speech utterances are randomly concatenated into a single *stream*. This stream is segmented into fixed-duration chunks that capture short-term context. Given the desired number N of classes, N chunks are taken at a time, and data points in the n-th chunk are assigned to class label n where $n \in \{1, 2, ..., N\}$. The process is repeated until all data points have been assigned. A segment-based clustering algorithm [28, 29] is applied to group similar chunks together and update the class labels until convergence. We then train a DNN to discriminate the obtained labels. The frame-level output of a hidden layer is projected into a low dimensional space using PCA to obtain sTCL-BN features.

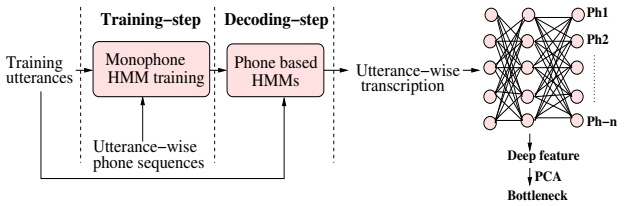

Fig. 1. Phone-BN feature extraction system.

3. EXPERIMENTAL SETUP

3.1. Dataset description

The audio data for the SdSV challenge was created from the DeepMine dataset collected by crowdsourcing [30, 31]. Task 1 involves speech files from 963 speakers as *in-domain* audio data. There is no development set which could be used for parameter tuning and optimization. We have created a development set by randomly choosing a subset of 63 speakers from this in-domain data. Similarly to the evaluation set, we enroll each speaker with three utterances for phrase-specific targets. If adequate data for a speaker-phrase combination is not available, we disregard that target model. The remaining sentences are used for test. We obtain a total of 519 target models and 4,810 speech utterances for test in the development set. The number of trials is summarized in Table 2. Similarly to the evaluation set, there is no cross-lingual trial. We have also discarded same-gender trials by clustering i-vectors for the 63 speakers (averaged over all utterances) into two *pseudo-gender* classes and keeping the trials for which the speakers in the enrollment model and the test utterance fall into the same class.

Table 2. Number of trials per condition in the development set.

TC	TW	IC	IW	Total
4,810	19,236	119,737	478,946	622,729

The remaining in-domain data from 900 speakers consisting of 94,661 utterances are considered for system development in addition to the other permitted audio data, such as VoxCeleb and LibriSpeech. The evaluation set consists of 12,404 enrollment models, 69,542 utterances, and 8,306,700 trials.

3.2. Dataset and parameters for system training

We refer the seven subsystems in the UIAI entry as S1–S7.

S1 is the UV system and it uses *mel-frequency cepstral coefficient* (MFCC) features. We extract 20-dimensional static MFCCs, apply a RASTA filter [32], and compute deltas and double-deltas to create 60-dimensional features. Utterance-level *cepstral mean and variance normalization* (CMVN)

is applied after discarding non-speech frames using energy-based activity detection (SAD). A 512-component UBM is trained on in-domain training data. The **T**-matrix is estimated with 600 factors on the same audio data. The extracted i-vectors are projected to 9 dimensions using LDA based on phrase labels.

S2 is the x-vector-PLDA ASV system. The x-vector extractor is trained on YouTube audio data from VoxCeleb-1 [33] and VoxCeleb-2 [34]. Recordings from the same YouTube video source are concatenated together. The scoring back-end (PLDA, centering, whitening, AS-norm) is trained on the in-domain data. For PLDA, the training labels are pairs of speaker and phrase IDs. Both in-domain and VoxCeleb data are augmented 5-fold using Kaldi's [35] augmentation recipes which include reverberation and additional noise, babble, or music. The input features for the embedding extractor are 60-dimensional, cepstral mean normalized MFCCs extracted with Kaldi. Kaldi's energy based VAD is applied to remove non-speech frames.

S3–5 are the GMM-UBM based ASV systems. Ten phrase-specific 512-component GMMs are trained on the in-domain data and merged into a 5120-component UBM. The target speaker models are created by MAP adaptation with a relevance factor of 3. We use three different acoustic front-ends. System **S3** is based on 60-dimensional MFCCs including deltas and double-deltas. System **S4** uses linear-frequency cepstral coefficients (LFCCs) with the same dimension. System **S5** uses 66-dimensional *overlapped block transform coefficients* (OBTCs) computed with two blocks of sizes 9 and 13 [36]. The pre- and post-processing stages are identical for the three feature sets. We use 20 mel filters and retain the energy coefficients. The features are processed with RASTA filtering and utterance-level CMVN. We do not apply SAD since this degraded performance.

S6 is the PBM-based joint verification system with Phone-BN features. The DNN feature extractor consists of 7 fully-connected layers with 1024 neurons and *sigmoid* activation. Its inputs are 57-dimensional RASTA-filtered MFCCs including deltas and double-deltas with a context of 11 frames. The number of outputs is 42. The DNN is trained on the in-domain data using CNTK [37]. We compute BN features by projecting the output of the second hidden layer on each time frame into a 57-dimensional vector using PCA. The PCA matrix is computed on 10,000 randomly selected utterances from the training part of LibriSpeech. We use the open-source *robust voice activity detector* (rVAD) [38] to discard non-speech frames from the enrollment and test utterances before applying utterance-level CMVN. The HMM used for extracting the phone labels does not discard non-speech frames, but it uses utterance-level CMVN. We use HTK [27] to build the HMM system. The Phone-BN features are then used with the PBM system where a gender-independent GMM-UBM with 2048 Gaussian components is trained using $60,000$ utterances from LibriSpeech. We create the target models using MAP adaptation with 3 iterations. We adapt only the mean vectors with a relevance factor of 10.

S7 is the PBM-based joint verification system with sTCL features. The DNN feature extractor uses the same input features as **S6**. We use rVAD [38] to discard non-speech frames, and we consider $N = 10$ unsupervised classes as in our previous study [28]. The DNN and the PBMs are trained on the same audio data using the same hyper-parameters as **S6**.

3.3. Performance evaluation

The primary metric for the challenge is the *normalized minimum detection cost function* (minDCF) [4]. The EER is also reported. These are computed by treating TC trials as genuine and by pooling the remaining three as impostor. We computed those metrics on the development set using our own scoring script while the challenge organizers computed performance on the evaluation set. We also report the performance for the three sub-conditions and their average values on the development set.

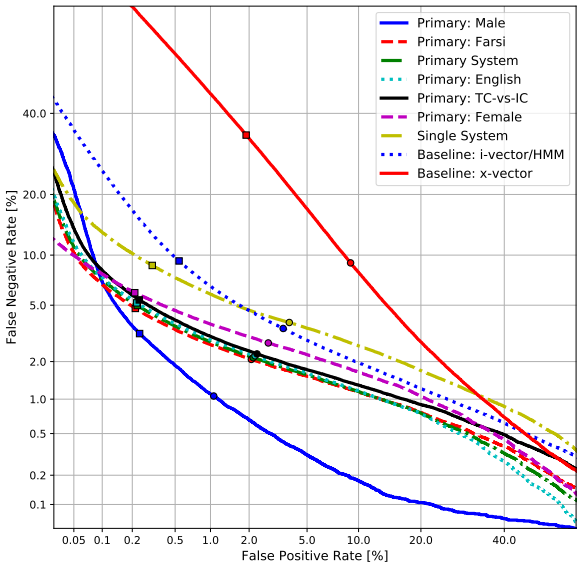

Fig. 2. DET plots of the UIAI systems along with baselines.

4. RESULTS

The ASV performance of the individual subsystems on the development set is shown in Table 3. The first row shows the ASV result of the UV system **S1** which performs well in wrong pass-phrase conditions (TW and IW). However, the performance for the IC condition is random with about 50% EER as this system does not use speaker information. The next row shows the performance for the x-vector-based ASV system **S2**. Although it yields promising performance in the

Table 3. Results (EER in % / minDCF) in the development set for the individual subsystems included in the UIAI primary submission.

ID	Method	Short-term features	Task	TW	IC	IW	Pooled	Avg.
S1	i-vector	MFCC	UV	0.08 / 0.005	51.64 / 1.00	0.08 / 0.005	18.29 / 1.00	17.27 / 0.337
S2	x-vector	MFCC	ASV	9.11 / 0.575	0.82 / 0.041	0.12 / 0.004	0.93 / 0.085	3.35 / 0.207
S3	GMM-UBM	MFCC	ASV	8.35 / 0.431	0.91 / 0.040	0.60 / 0.019	1.36 / 0.090	3.29 / 0.164
S4	GMM-UBM	LFCC	ASV	10.69 / 0.535	1.27 / 0.057	0.77 / 0.028	1.68 / 0.116	4.24 / 0.206
S5	GMM-UBM	OBTC	ASV	7.82 / 0.405	0.85 / 0.035	0.60 / 0.016	1.22 / 0.085	3.09 / 0.152
S6	PBM	Phone-BN	Both	0.07 / 0.008	1.31 / 0.051	0.01 / 0.001	0.81 / 0.029	0.46 / 0.020
S7	PBM	sTCL-BN	Both	0.14 / 0.011	1.74 / 0.062	0.01 / 0.001	1.07 / 0.038	0.63 / 0.025

Table 4. Results (EER in % / minDCF) of single and primary systems in the development and evaluation sets. The x-vector based baseline achieves 9.05% EER and 0.529 minDCF while the i-vector/HMM baseline achieves 3.49% EER and 0.146 minDCF.

System	TW	IC	IW	Pooled	Avg.	Eval set. (Pooled)
Single (**S6**)	0.07 / 0.008	1.31 / 0.051	0.01 / 0.001	0.81 / 0.029	0.46 / 0.020	3.83 / 0.118
Primary (fusion **S1-S7**)	0.06 / 0.006	0.32 / 0.015	0.00 / 0.000	0.17 / 0.007	0.13 / 0.007	2.14 / 0.072

IC and IW conditions, it performs poorly in the TW condition. Similarly, the GMM-UBM systems perform relatively well in the IC and IW conditions but they fail in the TW condition. Our results also indicate that the GMM-UBM systems give competitive performance compared to the x-vector system. The short-term OBTC features outperform MFCCs in all cases. Out of the two PBM-based methods, the one based on Phone-BN features performs consistently better and both of these methods outperform other subsystems in terms of average EER and minDCF. **S6** achieves the lowest average and pooled EERs as well as minDCFs. For this reason, we select it as the single system for the challenge.

We build the primary system submitted to the challenge by combining the modules for UV and ASV. We fuse **S1**, **S6** and **S7** for the UV task and all the subsystems except **S1** for the ASV task. The subsystems for each task are combined by linear score weighting where the weights are optimized on the development set by linear search. The UV and ASV system are then combined by cascade fusion as described earlier. The trials with wrong pass-phrases as detected by the fused UV system are assigned an ASV score of -100, while the trials with correctly detected pass-phrases are retained with fused ASV score.

Table 4 summarizes the results achieved by the single system and the primary system on the development and evaluation sets. The single system outperforms the two challenge baselines, especially in terms of minDCF. Figure 2 shows the DET plots of the submitted systems along with the baselines.

5. CONCLUSIONS

We have described the UIAI systems submitted to the SdSV challenge 2020 for text-dependent ASV. The systems developed are a fusion of different subsystems using various front-ends and back-ends. To deal with the pass-phrase verification problem, we combined the UV system with ASV in a cascade mode. Our development set created with a subset of limited in-domain data generalized well to the evaluation set by estimating suitable fusion parameters and by demonstrating systematic improvement. However, there was a substantial performance gap between the development and evaluation set possibly due to the large number of speakers and presence of unknown noises in the evaluation set.

Systems studied for the challenge submissions were developed independently by four different sites using various features and classifiers. This work could be extended towards a more systematic comparison of front-end acoustic features and different back-end classifiers. For instance, our phone-bottleneck features could be studied with x-vector system adopted for this work.

6. ACKNOWLEDGMENT

Experiments presented in this paper were partially carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`). This work has been partially sponsored by Academy of Finland (proj. no. 309629).

7. REFERENCES

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[2] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[3] A. Poddar, M. Sahidullah, and G. Saha, "Speaker verification with short utterances: a review of challenges, trends and opportunities," *IET Biometrics*, vol. 7, no. 2, pp. 91–101, 2017.

[4] H. Zeinali, K.-A. Lee, J. Alam, and L. Burget, "Short-duration speaker verification (SdSV) challenge 2020: the challenge evaluation plan," https://sdsvc.github.io/assets/SdSV_Challenge_Evaluation_Plan.pdf, accessed: 2020-04-20.

[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 *Proc. INTERSPEECH*, 2015, pp. 2996–3000.

[7] A. K. Sarkar and Z.-H. Tan, "Text dependent speaker verification using un-supervised HMM-UBM and temporal GMM-UBM," in *Proc. INTERSPEECH*, 2016, pp. 425–429.

[8] 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.

[9] 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 *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2016, pp. 24–30.

[10] S. Dey, P. Motlicek, S. Madikeri, and M. Ferras, "Template-matching for text-dependent speaker verification," *Speech Communication*, vol. 88, pp. 96–105, 2017.

[11] Q. He, G. W. Wornell, and W. Ma, "A low-power text-dependent speaker verification system with narrow-band feature pre-selection and weighted dynamic time warping," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2016, pp. 1–8.

[12] A. Sarkar and Z.-H. Tan, "Incorporating pass-phrase dependent background models for text-dependent speaker verification," *Computer Speech & Language*, vol. 47, pp. 259–271, 2018.

[13] T. Kinnunen, M. Sahidullah, I. Kukanov, H. Delgado, M. Todisco, A. K. Sarkar, N. B. Thomsen, V. Hautamäki, N. Evans, and Z.-H. Tan, "Utterance verification for text-dependent speaker recognition: A comparative assessment using the RedDots corpus," in *Proc. INTERSPEECH*, 2016, pp. 430–434.

[14] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," in *Proc. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 726–733.

[15] S. Wang, Y. Qian, and K. Yu, "What does the speaker embedding encode?" in *Proc. INTERSPEECH*, 2017, pp. 1497–1501.

[16] 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, 2010.

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

[18] D. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[21] H. Zhang, J. Xue, and K. Dana, "Deep TEN: Texture encoding network," in *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 708–717.

[22] W. Cai, Z. Cai, X. Zhang, X. Wang, and M. Li, "A novel learnable dictionary encoding layer for end-to-end language identification," in *Proc. 2018 IEEE International*

Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2018, pp. 5189–5193.

[23] V. Vestman, K. A. Lee, and T. H. Kinnunen, "Neural i-vectors," *arXiv preprint arXiv:2004.01559*, 2020.

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

[25] S. Cumani, P. D. Batzu, D. Colibro, C. Vair, P. Laface, and V. Vasilakakis, "Comparison of speaker recognition approaches for real applications," in *Proc. INTERSPEECH*, 2011, pp. 2365–2368.

[26] M. Sahidullah and T. Kinnunen, "Local spectral variability features for speaker verification," *Digital Signal Processing*, vol. 50, pp. 1–11, 2016.

[27] S. Young *et al.*, *The HTK Book (for version 3.4)*. Cambridge University Engineering Department, 2009.

[28] A. K. Sarkar, Z.-H. Tan, H. Tang, S. Shon, and J. R. Glass, "Time-contrastive learning based deep bottleneck features for text-dependent speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1267–1279, 2019.

[29] A. Sarkar and Z.-H. Tan, "Time-contrastive learning based DNN bottleneck features for text-dependent speaker verification," in *Proc. NIPS Time Series Workshop*, 2017.

[30] 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: The Speaker and Language Recognition Workshop*, 2018, pp. 386–392.

[31] 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. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 397–402.

[32] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

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

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

[35] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proc. 2011 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.

[36] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition," *Speech Communication*, vol. 54, no. 4, pp. 543–565, 2012.

[37] A. Agarwal *et al.*, "An introduction to computational networks and the computational network toolkit," *Microsoft Technical Report MSR-TR-2014-112*, 2016.

[38] Z.-H. Tan, A. K. Sarkar, and N. Dehak, "rVAD: An unsupervised segment-based robust voice activity detection method," *Computer Speech & Language*, vol. 59, pp. 1–21, 2020.