INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Analysis of BUT Submission in Far-Field Scenarios of VOiCES 2019 Challenge

Pavel Matějka, Oldřich Plchot, Hossein Zeinali, Ladislav Mošner, Anna Silnova, Lukáš Burget, Ondřej Novotný, Ondřej Glembek

Brno University of Technology, IT4I Center of Excellence, Brno, Czechia
{matejkap,iplchot,zeinali,burget,...}@fit.vutbr.cz

Abstract

This paper is a post-evaluation analysis of our efforts in VOiCES 2019 Speaker Recognition challenge. All systems in the fixed condition are based on x-vectors with different features and DNN topologies. The single best system reaches minDCF of 0.38 (5.25% EER) and a fusion of 3 systems yields minDCF of 0.34 (4.87% EER). We also analyze how speaker verification (SV) systems evolved in last few years and show results also on SITW 2016 Challenge. EER on the core-core condition of the SITW 2016 challenge dropped from 5.85% to 1.65% for system fusions submitted for SITW 2016 and VOiCES 2019, respectively. The less restrictive open condition allowed us to use external data for PLDA adaptation and achieve additional small performance improvement. In our submission to open condition, we used three x-vector systems and also one system based on i-vectors.

1. Introduction

Text-independent speaker verification (SV) field has already embraced Deep neural networks (DNN) for modeling in every stage and approaches such as end-to-end modelling and DNN embeddings became a new state-of-the-art. Domain adaptation and especially the need for a large amount of training data are still a challenge and unlike with generative models, we see significant gains from data augmentation, simulation and other techniques designed to overcome lack of training data.

We present an analysis of a SV system based on DNN embeddings (x-vectors) [1] and i-vectors [2] for far-field data in VOiCES 2019 challenge [3]. We also analyze the evolution of the SV systems submitted to Speakers In The Wild (SITW) 2016 challenge [4] till today. We show the differences in performance of the previous state-of-the-art based on i-vectors and current x-vector systems and at the same time we analyze the impact of different sampling frequency, training data and feature extraction on performance of the SV system for far-field data represented in SITW and VOiCES benchmarks.

Since our submissions back in 2016 for the SITW challenge and now for VOiCES were both very successful and can be considered state-of-the-art at the time, we are in a position to claim that in the last three years, there has been an enormous improvement in the performance of SV systems, especially for mostly English and far-field data. This progress was possible not only because of the new techniques like x-vectors, but thanks to the availability of large training sets like the VOXCELEB databases [5, 6] which suits these new approaches very well.

The main objective of this paper is not only to provide a description of our submission to the VOiCES challenge but also to provide an analysis of the evolution of the SV state-of-the-art in last few years.

2. Experimental Setup

2.1. Training data, Augmentations

We used VOXCELEB 1 and 2 datasets spanning 7146 speakers spread over 166 thousand sessions (distributed in 1.2 million speech segments totalling 450 hours of speech. For training both i-vector extractor and probabilistic linear discriminant analysis (PLDA), we concatenated all segments which belong to a single session. For training the x-vector DNN, we used original speech segments together with their augmentations. The augmentation process was based on the Kaldi recipe[1] and it resulted in additional 5 million segments belonging to following categories:

- Reverberated using RIRs[2]
- Augmented with Musan[3] noise
- Augmented with Musan music
- Augmented with babel noise obtained from Musan US-GOV speech part and Voxceleb 2 test part [4]

2.1.1. Retransmitted NIST SRE10 close talk data

In order to adapt our systems towards the far-field microphone data, we made use of our dataset of retransmitted audio [7] in the open condition track. Speaker verification part of this database has been benchmarked in [8]. A subset [5] of NIST 2010 Speaker Recognition evaluations (SRE) dataset was replayed by Adam audio A7X studio monitor in numerous rooms and acoustic conditions. In each room, multiple speaker positions were considered – sitting speaker, standing speaker and non-standard position (pointed to the ceiling, lying on the floor etc.). In addition to naturally occurring noise such as AC, vents, or common street noise coming through windows, noise source (radio receiver) was present in some sessions.

The corrupted audio was always simultaneously recorded by 31 microphones placed within the rooms. Synchronicity was governed by proprietary recording hardware.

The original dataset consists of 932 utterances with 30sec durations [6]. There are 459 recordings from 150 female speakers and 473 recordings from 150 male speakers. The whole set was retransmitted in 5 rooms. Changes of the loudspeaker positions in some of the rooms resulted in 9 recording sessions.

[1] https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v2

[2] http://www.openslr.org/resources/28/rirs_noises.zip

[3] http://www.openslr.org/17/

[4] We could not use the whole MUSAN for babble noise, because it uses LibriSpeech and this was against the challenge rules.

[5] We used mainly telephone recordings recorded over close talk microphones

[6] The original files have duration of 5 or 3 minutes, but we take only 30 sec chunks to limit overall retransmission time.

http://dx.doi.org/10.21437/Interspeech.2019-2471

2.2. Development and Evaluation data

2.2.1. SITW Challenge

SITW database (for more detailed description see [4]) is a large collection of real data exhibiting speech from individuals recorded in a wide array of challenging acoustic and environmental conditions. SITW include multi-speaker audio from both professionally edited interviews e.g. red carpet interviews, question and answer session in an auditorium etc., as well as more casual conversational multi-speaker audio in which backchannel, laughter, and overlapping is observed. Each individual also has raw, unedited camcorder or cellphone footage in which they speak, potentially with other speakers. All audio files do not contain any artificially added noise, reverberation or other artifacts. The audio of SITW was extracted from open-source media. Evaluation data consists of 180 speakers (2883 audio files). We report only results on core-core condition where audio files contain a continuous speech segment from a single speaker. The amount of enrollment speech is between 6-240 seconds.

2.2.2. VOiCES Challenge

The VOiCES from a distance challenge[3] focus on benchmarking and further improving state-of-the-art technologies in the area of speaker recognition and automatic speech recognition (ASR) for far-field speech. The data from the challenge is based on the recently released corpus Voices Obscured in Complex Environmental Settings (VOiCES)[9], where noisy speech was recorded in real reverberant rooms with multiple microphones. Noise sources include babble, music, or television. The database uses LibriSpeech [10] as source data for retransmission. Two sets were released for benchmarking the systems:

- Development data with 196 speakers, 15904 files (256 enrollment, 15648 test) providing 20k target and 3.98M non-target trials.
- Evaluation data with 100 speakers, 11392 files (328 enrollment, 11069 test) providing 36k target and 3.57M non-target trials.

2.3. Input features

We use different features for several systems with following settings:

- **Kaldi MFCC** - 16kHz, frequency limits 20-7600Hz, 25ms frame length, 40 filter banks, 30 coefficients + energy
- **HTK MFCC** - 16kHz, frequency limits 0-8kHz, 25ms frame length, 30 filter banks, 24 coefficients + energy [11]
- **Kaldi PLP** - 16kHz, frequency limits 20-7600Hz, 25ms frame length, 40 filter banks, 30 coefficients
- **Kaldi FBank** - 16kHz, frequency limits 20-7600Hz, 25ms frame length, 40 filter banks
- **SBN** - 8kHz, 80 dimensional stack bottleneck features (SBN) trained on Fisher English, more details in [12].

Kaldi MFCCs, PLPs and FBanks are subjected to short time mean normalization over 3sec window. For the HTK MFCC, we apply also short time variance normalization.

2.4. Voice Activity Detection

We used two VAD approaches:

VAD-Energy is an energy-based VAD from Kaldi SRE16 recipe.

VAD-NN The NN which produces per-frame posterior probabilities for speech and non-speech classes that are later post-processed to create continuous speech segments was trained on the 8kHz Fisher English [13].

3. i-vector Systems

The system is based on gender independent i-vectors [2, 14]. HTK MFCCs with deltas and double deltas together with SBN feature vectors were extracted from recordings (SBNs were extracted from audio downsampled to 8kHz). Final feature vector is a concatenation of both as they proved to perform very well in NIST SRE [15]. This system uses VAD-NN. Universal background model (UBM) contained 2048 Gaussian components and the i-vector subspace dimensionality was set to 600. UBM, i-vector extractor and PLDA were trained only on clean data.

For the purpose of PLDA training we preprocessed all training, enrollment and test data by the dereverberation system based on single-channel weighted prediction error (WPE) [16] to suppress effects of room acoustic conditions.

4. x-vector Systems

All x-vectors used VAD-Energy from Kaldi SRE16 recipe[7]. The systems were trained with Kaldi toolkit [18] using SRE16 recipe with modifications described below:

- Using different feature sets (MFCC, PLP, FBANK)
- Training networks with 9 epochs (instead of 3). We did not see any considerable difference with 12 epochs.
- Using modified example generation - we used 200 frames in all training segments instead of randomizing it between 200-400 frames. We also have changed generation of the training examples so that it is not random and uses almost all available speech from all training speakers.
- The x-vector DNN was trained on 1.2 million speech segments from 7146 speakers plus additional 5 million segments obtained with data augmentation. We generated around 700 Kaldi archives such that each of them contains exactly 15 training examples from each speaker (i.e. around 107K examples in each archive).
- The architecture of the network for x-vector extraction is shown in Table 1. There are 2 topologies - "standard" and "BIG".

5. Backend

5.1. Heavy-tailed PLDA

Only our final i-vector system for open condition used HT-PLDA backend [19]. It was trained on concatenated audio files from VOXCELEB 1 and 2.Length normalization, centering, LDA, reducing dimensionality of vectors to 300, followed by another length normalization were applied to all i-vectors. All i-vectors were centered using the mean computed on training data. We fixed the size of the speaker subspace to 200. Degrees of freedom parameter was set to infinity at the training time and to 2 at scoring time. Finally, we performed adaptive score normalization as described in Section 5.4. We did not use any augmented data for HT-PLDA training on i-vectors.

5.2. Gaussian PLDA

For all x-vector based systems as well as for most i-vector based systems we trained Gaussian PLDA backend. As in the case of

[7] We did not find a significant impact on performance when using different VAD within the x-vector paradigm and it seems that simple VAD from Kaldi performs very well for x-vectors in various channel conditions.

Table 1: *x-vector topology proposed in [17]. K in the first layer indicates different feature dimensionalities, T is the number of training segment frames and N in the last row is the number of speakers.*

Layer	Standard DNN		BIG DNN	
	Layer context	**(Input) × output**	**Layer context**	**(Input) × output**
frame1	$[t-2, t-1, t, t+1, t+2]$	(5 × K) × 512	$[t-2, t-1, t, t+1, t+2]$	(5 × K) × 1024
frame2	$[t]$	512 × 512	$[t]$	1024 × 1024
frame3	$[t-2, t, t+2]$	(3 × 512) × 512	$[t-4, t-2, t, t+2, t+4]$	(5 × 1024) × 1024
frame4	$[t]$	512 × 512	$[t]$	1024 × 1024
frame5	$[t-3, t, t+3]$	(3 × 512) × 512	$[t-3, t, t+3]$	(3 × 1024) × 1024
frame6	$[t]$	512 × 512	$[t]$	1024 × 1024
frame7	$[t-4, t, t+4]$	(3 × 512) × 512	$[t-4, t, t+4]$	(3 × 1024) × 1024
frame8	$[t]$	512 × 512	$[t]$	1024 × 1024
frame9	$[t]$	512 × 1500	$[t]$	1024 × 2000
stats pooling	$[0, T]$	1500 × 3000	$[0, T]$	2000 × 4000
segment1	$[0, T]$	3000 × 512	$[0, T]$	4000 × 512
segment2	$[0, T]$	512 × 512	$[0, T]$	512 × 512
softmax	$[0, T]$	512 × N	$[0, T]$	512 × N

HT-PLDA, we used concatenated data from VoxCeleb 1 and 2 for training. In this case, we train the backend only on x-vectors extracted from the original utterances augmented with reverberation and noise. X-vectors extracted from the non-augmented files were not used for backend training. Centering, LDA dimensonality reduction to 250 dimensions followed by length normalization was applied to x-vectors. All data were centered using the training data mean. Speaker and channel subspace size was set to 250 (i.e full rank). Same as in the case of HT-PLDA, we applied adaptive score normalization described in Section 5.4.

5.3. Adaptation (ADAPT)

For open condition, we used 280k files of BUT retransmitted data (see Section 2.1.1) to perform domain adaptation by model interpolation. That is, we train smaller G-PLDA model on retransmitted data, size of both speaker and channel subspaces was fixed to 150. The final adapted model is derived from the two G-PLDA models so that the modeled within- and across-speaker covariance matrices are a weighted combination of the covariance matrices from the constituent models. Similarly, the model means are also interpolated. Interpolation weights are set to 0.6 for the original model and 0.4 for the adaptation one. The systems which use this adaptation are denoted ADAPT in the Table 2.

5.4. Score normalization

We used adaptive symmetric score normalization (adapt S-norm) which computes an average of normalized scores from Z-norm and T-norm [14, 20]. In adaptive version [20, 21, 22], only part of the cohort is selected to compute mean and variance for normalization. Usually X top scoring or most similar files are selected, and we set X to be 400 for all experiments. The cohort is created from PLDA training data and consists of approximately 15k files (two files per speaker).

5.5. Calibration and Fusion

Each system provided log-likelihood ratio scores that could be subjected to score normalization. These scores were first pre-calibrated and then passed into the fusion. The output of the fusion was then again re-calibrated. Calibration and fusion was trained on the labeled VOiCES development data [3, 9] by the means of logistic regression optimizing the cross-entropy between the hypothesized and true labels. The parameters optimized during the fusion were single scalar offset and the scalar combination of system weights.

6. Results and Analysis

In this section we provide not only an analysis of our final submission, but we also look at various system designs and architectures that represent the evolution in SV state-of-the-art since 2016 (at least for wideband distant microphone data). You can view our analysis as a story which has 3 acts - telephone vs. youtube/microphone data, i-vector vs. x-vector architecture and 8kHz vs 16kHz bandwidth. All the results are in the Table 2. First part of the table (line 1 to 6) shows i-vector based systems trained on narrowband data. We start with a "conventional" system trained on telephone data from NIST SRE evaluations which has 20.4% EER on the VOiCES evaluation set. Systems 2 to 5 in the Table show effect of using different training data for i-vector extractor and PLDA. We are distinguishing between telephone NIST data and VOXCELEB data which are closer to the target domains of VOiCES and SITW challenges. The best results are obtained when both i-vector extractor and PLDA, are trained only on VOXCELEB data (system 5). In system 6, we can observe an effect of using MFCCs concatenated with SBNs which is still our favorite approach to consider when using i-vectors. We see an improvement on the SITW and VOiCES dev, but almost no improvement on VOiCES eval.

Second part of the Table 2 shows 16kHz MFCC (rows 7-9) and PLP systems (rows 10-12) which performs about the same. Comparing systems 5 and 7 or 6 and 8 we can see a very nice improvement when switching from 8kHz to 16kHz. We run WPE as preprocessig of the PLDA training data, enrollment and test data for systems 9 (MFCC) and 12 (PLP) to deal with the reverberated data in the challenge. By comparing systems 8 and 9 for MFCC or 11 and 12 for PLP we see that there is a small but consistent improvement on all conditions from using WPE.

Third part of the Table 2 brings x-vectors into the game. At first we can see a dramatic 40% relative improvement when comparing i-vector (system 5) to 8kHz x-vector (system 13). Switching from 8kHz x-vector (13) to 16kHz x-vector (14) gives us another almost 30% relative improvement.

Comparing systems 14 and 15 shows the impact of changing the x-vector DNN topology and its size which is shown in the right part of the Table 1. Bigger DNN provides 5% relative improvement over the standard size in system 14. Rows 16 and 17 are the same x-vector architecture as 14, except the different input features - FBANK and PLP, respectively. System based on FBANK features has a slight edge over the PLPs and MFCCs.

Next block of the Table 2 (systems 18-20) shows how adaptive score normalization (s-norm) helps. There is a consistent

Table 2: *Analysis of the systems on SITW and VOiCES challenges.*

#	System Name/Configuration	SITW core-core		VOiCES dev		VOiCES evl	
		MinDCF	EER	MinDCF	EER	MinDCF	EER
1	8kHz MFCC ivec&PLDA(NIST)	0.705	11.2	0.910	14.5	0.975	20.4
2	8kHz MFCC ivec(NIST), PLDA(NIST+VOXCELEB)	0.571	7.33	0.895	12.4	0.957	16.8
3	8kHz MFCC ivec(NIST)&PLDA(VOXCELEB)	0.543	6.32	0.898	11.7	0.944	15.6
4	8kHz MFCC ivec&PLDA(NIST+VOXCELEB)	0.560	7.21	0.901	11.7	0.952	16.7
5	8kHz MFCC ivec&PLDA(VOXCELEB)	0.524	5.87	0.870	10.7	0.928	15.2
6	8kHz MFCC+SBN ivec&PLDA(VOXCELEB)	0.481	4.99	0.836	9.60	0.917	15.2
7	16kHz MFCC ivec&PLDA(VOXCELEB)	0.371	4.40	0.587	6.18	0.787	11.5
8	16kHz MFCC+SBN ivec&PLDA(VOXCELEB)	0.334	3.62	0.516	5.65	0.744	12.0
9	16kHz WPE MFCC+SBN ivec&PLDA (VOXCELEB)	0.330	3.36	0.475	5.03	0.703	11.2
10	16kHz PLP ivec&PLDA(VOXCELEB)	0.360	3.96	0.583	6.81	0.782	12.0
11	16kHz PLP+SBN ivec&PLDA(VOXCELEB)	0.336	3.50	0.532	5.60	0.761	11.9
12	16kHz WPE PLP+SBN ivec&PLDA (VOXCELEB)	0.325	3.38	0.492	5.00	0.699	11.2
13	8kHz MFCC xvec&PLDA (VOXCELEB)	0.298	2.73	0.494	4.68	0.749	9.87
14	16kHz MFCC xvec&PLDA (VOXCELEB)	0.213	1.90	0.260	2.04	0.481	6.04
15	16kHz MFCC xvecBIG&PLDA (VOXCELEB)	0.194	1.71	0.219	1.76	0.435	5.72
16	16kHz FBANK xvec&PLDA (VOXCELEB)	0.203	1.70	0.191	1.62	0.417	5.65
17	16kHz PLP xvec&PLDA (VOXCELEB)	0.200	1.85	0.206	1.81	0.438	5.79
18	16kHz MFCC xvecBIG&PLDA (VOXCELEB) + snorm	0.177	1.77	0.163	1.29	0.399	5.31
19	16kHz FBANK xvec&PLDA (VOXCELEB) + snorm	0.188	1.80	0.141	1.20	0.376	5.24
20	16kHz PLP xvec&PLDA (VOXCELEB) + snorm	0.191	1.92	0.163	1.44	0.380	5.36
21	16kHz FBANK xvec&PLDA (VOXCELEB) + ADAPT + snorm	0.202	1.99	0.146	1.13	0.365	4.90
22	16kHz PLP xvec&PLDA (VOXCELEB) + ADAPT + snorm	0.195	2.11	0.157	1.31	0.365	4.99
23	VOiCES2019 Fusion Fixed = 18+19+20 - submission	0.174	1.65	0.122	1.04	0.338	4.86
24	VOiCES2019 Fusion Fixed - no snorm	0.191	1.50	0.171	1.44	0.390	5.27
25	VOiCES2019 Fusion Open = 12+18+21+22 - submission	0.174	1.73	0.120	1.00	0.322	4.69
26	SITW2016 8kHz MFCC+SBN ivec&PLDA(NIST) [23]	0.560	7.72	0.817	11.0	0.918	17.2
27	SITW2016 16kHz MFCC ivec&PLDA(NIST) [23]	0.713	9.34	0.767	9.21	0.892	16.2
28	SITW2016 16kHz PLP ivec&PLDA(NIST) [23]	0.688	9.22	0.795	10.0	0.892	16.3
29	SITW2016 - fusion 26+27+28 - submission BUT [23]	0.503	5.85	0.602	6.98	0.782	12.6

improvement over the systems without s-norm (15-17). Fusion of these three systems (18-20) form our primary submission (system 23) to the fixed condition. We have also run a posteevaluation fusion with the same systems without s-norm (15-17) which is show in row 24. There is a clear benefit of using s-norm for VOiCES challenge.

Fusion for the open condition is the row 25 and there is a negligible improvement over the fixed condition fusion in the row 23, despite having small gains with adaptation in individual systems (compare systems 19 and 21 for FBANK and 20 to 22 for PLP).

Last part of the Table 2 is about the system we developed for the SITW 2016 challenge. We took the same three systems and their fusion that we had back in 2016 and evaluated them on VOiCES dataset to see where we moved during the last 3 years. There are two main reasons why our current results are much better. First, from the technological point of view, it is the introduction of the x-vector model and the second reason is the availability of the VOXCELEB dataset that is large both in terms of amount of speakers and audio. It is also important to note that VOXCELEB contains mostly English microphone data which well matches the conditions of both VOiCES and SITW challenges and therefore brings substantial improvements.

7. Conclusion

BUT participated in both SITW and VOiCES challenges and therefore we are in a position to compare what happened in the 3 years time period in the main stream of the speaker verification state-of-the-art. Conveniently, the domains of the two challenges are very similar, which makes our analysis on both datasets relevant and interesting. Our submission to SITW back in 2016 reached 5.85% EER, while now the best fusion reached 1.65% which is an enormous improvement especially considering the fact that we did not specifically targeted performance on the SITW this time. When comparing the same systems on the VOiCES challenge, the performance is 12.6% EER and 4.86% EER for the system from 2016 and for the current system, respectively, which is more than 50% relative improvement. The improvement comes from the new technology based on DNN (x-vector) and also from releasing new data (VOXCELEB) which are close to the target domain.

8. Acknowledgements

The work was supported by Czech Ministry of Interior project No. VI20152020025 "DRAPAK", Google Faculty Research Award program, Czech Science Foundation under project No. GJ17-23870Y, Czech National Science Foundation (GACR) project "NEUREM3 No. 19-26934X, and by Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602". It was also supported by TACR project No. TJ01000208 "NOSICI", European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 748097, the Marie Sklodowska-Curie cofinanced by the South Moravian Region under grant agreement No. 665860, and by the U.S. DARPA LORELEI contract No. HR0011-15-C-0115. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

9. References

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

[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. PP, no. 99, pp. 1 –1, 2010.

[3] Mahesh Kumar Nandwana, Julien van Hout, Mitch McLaren, Aaron. Lawson, and María Auxiliadora Barrios, "The voices from a distance challenge 2019 evaluation plan," in *arXiv:1902.10828 [eess.AS]*, 2019.

[4] Diego Castan Aaron Lawson Mitchell McLaren, Luciana Ferrer, "The speakers in the wild (SITW) speaker recognition database," in *Interspeech 2016*, 2016.

[5] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, 2017, pp. 2616–2620.

[6] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.*, 2018, pp. 1086–1090.

[7] Igor Szöke, Miroslav Skácel, Ladislav Mošner, Jakub Paliesek, and Jan Černocký, "Building and Evaluation of a Real Room Impulse Response Dataset," *Under review for IEEE Journal of Selected Topics in Signal Processing*, 2019.

[8] Ladislav Mošner, Oldřich Plchot, Pavel Matějka, Ondřej Novotný, and Jan Černocký, "Dereverberation and Beamforming in Robust Far-Field Speaker Recognition," in *Proceedings of Interspeech 2018*. 2018, pp. 1334–1338, International Speech Communication Association.

[9] Colleen Richey, María Auxiliadora Barrios, Zeb Armstrong, Chris Bartels, Horacio Franco, Martin Graciarena, Aaron Lawson, Mahesh Kumar Nandwana, Allen R. Stauffer, Julien van Hout, Paul Gamble, Jeff Hetherly, Cory Stephenson, and Karl Ni, "Voices obscured in complex environmental settings (VOICES) corpus," in *ISCA INTERSPEECH 2018*, 2018.

[10] 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.

[11] S. Young et al., *The HTK Book*, University of Cambridge, 2005.

[12] Pavel Matějka et al., "Neural network bottleneck features for language identification," in *IEEE Odyssey: The Speaker and Language Recognition Workshop*, Joensu, Finland, 2014.

[13] Oldřich Plchot, Pavel Matějka, Ondřej Novotný, Sandro Cumani, Alicia Díez Lozano, Josef Slavíček, Mireia Sánchez Diez, František Grézl, Ondřej Glembek, Mounika Veera Kamsali, Anna Silnova, Lukáš Burget, Lucas Ondel, Santosh Kesiraju, and A. Johan Rohdin, "Analysis of but-pt submission for nist lre 2017," in *Proceedings of Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 47–53.

[14] P. Kenny, "Bayesian speaker verification with heavy–tailed priors," keynote presentation, Proc. of Odyssey 2010, June 2010.

[15] Pavel Matějka, Ondřej Glembek, Ondřej Novotný, Oldřich Plchot, František Grézl, Lukáš Burget, and Jan Černocký, "Analysis of dnn approaches to speaker identification," in *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2016*. 2016, IEEE Signal Processing Society.

[16] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

[17] David Snyder, Daniel Garcia-Romero, Gregory Sell, Alan McCree, Daniel Povey, and Sanjeev Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP*, 2019.

[18] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[19] Anna Silnova, Niko Brummer, Daniel Garcia-Romero, David Snyder, and Lukáš Burget, "Fast variational bayes for heavy-tailed plda applied to i-vectors and x-vectors," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.*, 2018.

[20] Pavel Matějka, Ondřej Novotný, Oldřich Plchot, Lukáš Burget, Mireia Sánchez Diez, and Jan Černocký, "Analysis of score normalization in multilingual speaker recognition," in *Proceedings of Interspeech 2017*. 2017, pp. 1567–1571, International Speech Communication Association.

[21] D. E. Sturim and Douglas A. Reynolds, "Speaker adaptive cohort selection for tnorm in text-independent speaker verification," in *ICASSP*, 2005, pp. 741–744.

[22] Yaniv Zigel and Moshe Wasserblat, "How to deal with multiple-targets in speaker identification systems?," in *Proceedings of the Speaker and Language Recognition Workshop (IEEE-Odyssey 2006)*, San Juan, Puerto Rico, June 2006.

[23] Ondřej Novotný, Pavel Matějka, Oldřich Plchot, Ondřej Glembek, Lukáš Burget, and Jan Černocký, "Analysis of speaker recognition systems in realistic scenarios of the sitw 2016 challenge," in *Proceedings of Interspeech 2016*, 2016, pp. 828–832.