INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Analysis of Speaker Recognition Systems in Realistic Scenarios of the SITW 2016 Challenge

Ondřej Novotný, Pavel Matějka, Oldřich Plchot, Ondřej Glembek, Lukáš Burget, and Jan "Honza" Černocký

Brno University of Technology, Speech@FIT and IT4I Center of Excellence

Abstract

In this paper, we summarize our efforts for the Speakers In The Wild (SITW) challenge, and we present our findings with this new dataset for speaker recognition. Apart from the standard comparison of different SRE systems, we analyze the use of diarization for dealing with audio segments containing multiple speakers, as in part of the newly introduced enrollment and test protocols, diarization is a necessary system component. Our state-of-the-art systems used in this work utilize both cepstral and DNN-based bottleneck features and are based on i-vectors followed by Probabilistic Linear Discriminant Analysis (PLDA) classifier and logistic regression calibration/fusion. We present both narrow-band (8 kHz) and wide-band (16 kHz) systems together with their fusions.

1. Introduction

The state-of-the-art in the text independent speaker recognition (SRE) has been fairly stable in the past years. Systems based on i-vectors [1] and PLDA [2] are still the architecture of choice for SRE systems in various acoustic conditions [3, 4]. Recently, we have seen an improvement from using DNNs that are trained to classify phoneme states [5, 6]. This DNN can be used at various places in the SRE system chain. In [7], we have extensively experimented with using DNNs for directly collecting sufficient statistics, using it as a source of frame alignment for Gaussian Mixture Models (GMM) or we used them as a feature extractor to provide bottleneck (BN) features [8, 9] that are best used in concatenation with standard Mel-frequency cepstral coefficients (MFCC). The role of the DNN as the feature extractor suits best to our needs as it provides very good and stable results and fits easily to the well-tested i-vector framework.

We have gladly accepted the opportunity to participate in the SITW challenge as it allows us to test state-of-the-art SRE systems in new channel domains and modify them in such a way that they can deal with multiple speakers being present in enrollment and test segments. The new data-collection protocol[1] that utilizes data from public sources brings new challenges and problems that are needed in order to advance the SRE technology towards additional use cases. One of the obvious properties of such dataset is the presence of variety of real noise and reverberation, which will test the robustness of current systems. Already mentioned possibility of multiple speakers speaking in audio segments lead to the introduction of a new enrollment and test protocol that is based on providing an annotation for only a small part of the enrollment segment. It is then up to the system (speaker diarization component) to search for additional data corresponding to the same speaker that is speaking in the annotated part of the enrollment segment and correctly segmenting the test segment between different possible speakers.

The sampling frequency of provided data is 16 kHZ and therefore we also developed wideband versions of our SRE systems. Because of the lack of suitable wideband data, the development of these systems was difficult and their performance low. In spite of their lower performance, we have seen a positive effect of fusing 16 kHz and 8 kHz systems.

This work was supported by the DARPA RATS Program under Contract No. HR0011-15-C-0038. The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. The work was also supported by Czech Ministry of Interior project No. VI20152020025 "DRAPAK"

[1] http://www.speech.sri.com/projects/sitw/

2. Data

2.1. Training Data

In 8 kHz system, we used PRISM dataset [10]. PRISM contains data from all NIST SREs, beginning with the year 2005 until 2010 [11]. Also, NIST 2004, Switchboard (Phase 1 and 2 and Cellphone phase 1 and 2) and Fisher (1 and 2) data are included in the dataset for training purpose. In total, PRISM dataset contains 16247 speakers and 86681 audio files.

We used 15602 audio files (1179 speakers) for UBM training and 108269 audio files (3585 speakers) for PLDA training. PLDA training set includes a portion of audio with artificially added crowd noise and short duration segments, which we created by extracting 20-160s long segments from PRISM.

In the 16 kHz version of the system, we used 16 kHz MIC recordings from AMI corpus (consist of 100 hours of meeting recordings) and NIST SRE 2010.

In total, we used 17664 audio files (17664 speakers) for UBM training and 21211 audio files (458 speakers) with artificially added crowd noise for PLDA training.

2.2. Development and Evaluation Data

SITW database (for more detailed description see [12]) is a large collection of real data exhibiting speech from individual across 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.) All audio files do not contain any artificially added noise, reverberation or other artifacts. The audio of SITW was extracted from open-source media.

Development data consist of 119 of actual speakers (1958 audio files). This set could be used without any restriction. Evaluation data consists of 180 speakers (2883 audio files).

The three enrollment conditions are defined as:

- **Core**: Audio files each containing a continuous speech segment from a single speaker. The amount of enrollment speech is between 6-240 seconds.

http://dx.doi.org/10.21437/Interspeech.2016-981

- **Assist**: Audio files which can contain one or more speakers and, at least, one POI (Point Of Interest) speaker. The audio from which POI speakers are enrolled in this condition is assisted.
- **AssistClean**: This condition is a subset of the *Assist* condition,

The two test conditions are defined as:

- **Core**: Audio files contains speech from a single speaker. The amount of speech per file is approximately 6-240 seconds.
- **Multi**: Audio files contains one or more speakers. This is a superset of the Core conditions. The amount of speech in each file vary from approximately 6 seconds to 10 minutes. When POI speaker is present in a file, the file contains at least 6 seconds of speech from that speaker.

3. System Components

3.1. Voice Activity Detection

Voice activity detection (VAD) was performed by the BUT Czech phoneme recognizer [13], dropping all frames that are labeled as silence or noise. The recognizer was trained on the Czech CTS data, but we have added noise with varying SNR to 30% of the database.

3.2. Diarization

Our speaker diarization was based on the Variational Bayes method described in [14, 15]. However, we generalize the method in our implementation by using an HMM instead of the simple mixture model when the modeling generation of segments (or even frames) from speakers. HMM limits the probability of switching between speakers when changing frames, which makes it possible to use the model on frame-by-frame basis without any need to iterate between 1) clustering speech segments and 2) re-segmentation (i.e. as it was done in the paper above).

We used 19 MFCC+Energy coefficients (without any normalization) as features for diarization, which we ran only on segments containing speech according to our VAD. We used 1024-component, diagonal covariance GMM UBM, and a factor loading matrix with 400 eigen-voices (JFA $\mathbf{V}$ matrix). The UBM and the $\mathbf{V}$ matrix were trained on the clean portion of the PRISM set [10]. We ran the diarization 5 times with different random initialization, and we picked the diarization output with the highest likelihood.

3.3. Feature Extraction

3.3.1. MFCC $8\,kHz$ and $16\,kHz$

These front-ends operate on standard Mel-Frequency Cepstrum Coefficients (MFCC). We implemented two versions, based on audio sampling frequency.

In 8kHz version, MFCCs were extracted using a 25ms Hamming window. We extract 19 MFCCs together with log-energy every 10 ms. MFCCs were augmented with delta and double delta coefficients calculated using a 5 frame window. Resulting 60-dimensional vectors are subjected to feature warping using the 3s sliding window before removing the silence.

In 16 kHz version of this front-end, we extracted 24 MFCCs from 30 filter-banks. The bandwidth was 0-8 kHz.

3.3.2. PLP $16\,kHz$

We also used Perceptual Linear Prediction (PLP) features. PLPs were extracted using 25 ms Hamming window. 24 PLP coefficients with normalized energy from 30 filter-banks with bandwidth 0-8 kHz were extracted every 10 ms.

3.3.3. SBN $8\,kHz$

Bottleneck Neural-Network (BN-NN) refers to such topology of a NN, where one of the hidden layers has significantly lower dimensionality than the surrounding layers. A bottleneck feature vector is generally understood as a by-product of forwarding a primary input feature vector through the BN-NN and reading off the vector of values at the bottleneck layer. We have used a cascade of two such NNs for our experiments. The output of the first network is *stacked* in time, defining context-dependent input features for the second NN, hence the term Stacked Bottleneck features.

The NN input features are 24 log Mel-scale filter bank outputs augmented with fundamental frequency features from 4 different f_0 estimators (Kaldi, Snack[2], and other two according to [16] and [17]). Together, we have 13 f_0 related features, see [18] for more details. The conversation-side based mean subtraction is applied on the whole feature vector. 11 frames of log filter bank outputs and fundamental frequency features are stacked together. Hamming window followed by DCT consisting of 0^{th} to 5^{th} base are applied on the time trajectory of each parameter resulting in $(24 + 13) \times 6 = 222$ coefficients on the first stage NN input.

The configuration for the first NN is $222 \times D_H \times D_H \times D_{BN} \times D_H \times K$, where K is the number of targets. The dimensionality of the bottleneck layer, D_{BN} was fixed to 80. This was shown as optimal in [19]. The dimensionality of other hidden layers was set to 1500. The bottleneck outputs from the first NN are sampled at times $t-10$, $t-5$, t, $t+5$ and $t+10$, where t is the index of the current frame. The resulting 400-dimensional features are inputs to the second stage NN with the same topology as first stage. The 80-dimensional bottleneck outputs from the second NN (referred as SBN) are taken as features for the conventional GMM/UBM i-vector based SID system.

3.4. I-vector Extraction

Both 8 kHz and 16 kHz systems are based on gender independent i-vectors [1, 20].

In 8 kHz version of the system, we used the feature-level fusion of MFCC and SBN, which proved to perform very well in NIST SRE [7]. Reverberated and artificially created noisy data from original PRISM set were not used. Another modification to the PRISM set was the use of short duration segments, which we created by extracting 20-160s long segments from our original training data.

We built two 16 kHz systems. The first uses MFCCs with 1024 diagonal Gaussian component GMM-UBM and 400-dimensional i-vector extractor. The second system is based on PLPs with 1024 diagonal Gaussian component GMM-UBM and 600-dimensional i-vector extractor.

Both systems were trained on the AMI and NIST SRE 2010 MIC data (only the ones distributed in 16kHz). UBM was trained on 17664 files collected from AMI and NIST SRE 2010 without 8min interview files. The i-vector extractor was trained on all 16 kHz data from both corpora.

[2] http://kaldi.sourceforge.net, www.speech.kth.se/snack/

3.5. Classifier

We used a standard flavor of Probabilistic Linear Discriminant Analysis (PLDA) with i-vectors. I-vectors were transformed by LDA into 200 dimensions and then we applied length normalization [1, 21]. We did not use any SITW dev data for training the PLDA. We trained one gender-independent PLDA per system.

In 8 kHz system, LDA and PLDA were trained also on the files where crowd noise was added at various levels of SNRs to the clean microphone data. Reverberated and artificially created noisy data from the original PRISM set were not used. The crowd noise samples were composed by summing recordings from the Fisher database.

In 16 kHz system, PLDA was trained only on NIST SRE 2010. We added crowd noise to 1200 files from NIST SRE data to make the system more robust. Crowd noise was created by summing audio files from the AMI corpus.

3.6. System Calibration and Fusion

We used only the SITW development set to train our calibration and fusion. We ran a separate calibration and fusion for each condition. Both calibration and fusion were trained as a logistic regression optimizing the cross-entropy on the development set. We employed the jack-knifing scheme to obtain valid results on SITW development data. We used the BOSARIS toolkit [22] to perform both steps.

4. Experiments and Discussion

4.1. Assisted Enrollment

Assisted enrollment tackles the issue of not having a single speaker in one speech file. In this condition, the enrollment is defined by an annotated region in the audio file (typically 5 seconds). However, the protocol allows to use the speech from the entire file. We use the diarization output to associate all speech frames with the annotated speaker.

Our diarization may output multiple speakers within the annotated region. We used the following procedure to find speech regions for the enrollment speaker:

1. Within the annotated region, identify speaker with the most detected speech, and let N_0 be the length of the associated speech.
2. Within the annotated region, identify all speakers which have more than 20 % of N_0 of detected speech (implicitly including speaker from step 1).
3. Within the whole recording, select speech associated with all speakers from step 2, and compute the enrollment i-vector.

An example of segments selection for i-vector extraction is shown in Fig. 1.

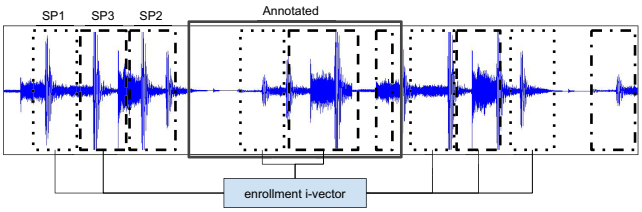

Figure 1: Post-diarization selection of segments for i-vector extraction in enrollment audio.

4.2. Multi-speaker Testing

We computed the i-vector for each speaker detected by the diarization, and then we scored each such i-vector against the i-vector representing the enrollment speaker. The maximum of all scores (log-likelihood ratios) was selected as the final score.

4.3. Performance Measures

The primary metric (DCF) for the SITW 2016 is based on detection cost function as defined in the NIST 2010 SRE [23] with the prior probability of the target trial set to 0.01. Costs for misses and false alarm are set to 1. We also report Equal Error Rates (EER) and $\mathrm{DCF}_{\mathrm{min}}$.

4.4. System Analysis

We designed three systems based on the components described in Sec. 3: 1) 16 kHz PLP system, 2) 16 kHz MFCC system, and 3) a system based on the feature-level fusion of 8 kHz MFCCs and SBN. The overall results are presented in Tab. 1.

Our further analysis is based on the MFCC+SBN system. Tab. 2 shows a brief comparison of the gender-dependent performance. We see that—in terms of EER—the system performance on the female trials is more than twice worse than on the male trials.

4.4.1. Diarization Analysis

The next system component subjected to analysis is diarization, because it is the key component in most of the condition in the SITW evaluation. In diarization analysis, we target on three experiments testing our system under different conditions. We use results from *assistclean-core* and *core-multi* condition to compare results.

In the first set of experiments, we compare different strategies to extracting the enrollment i-vectors in the *assistclean-core* condition. Tab. 3 presents the results. In the first case, we used the whole utterance for i-vector extraction regardless of the diarization output (marked as "no diarization" in the table). In the second case, we used only annotated part of the enrollment audio in i-vector extraction (marked as "annotated part"). In the third case, we used our diarization technique as described in Sec. 3.2 (marked as "diarization"). All of these experiments were also divided into groups based on the number of speakers in the enrollment audio. Based on the results, we can state that the diarization technique does not harm system in the case, where we have only one speaker in enroll audio. In all other groups, we see that it is beneficial to use diarization.

In the second set of experiments, we study the effect of the length of the annotated part of the enrollment audio in the *assistclean-core* condition with respect to the "diarization" case, as described in the previous set of experiments. In other words, it studies whether more annotated speech leads to better spotting the same speaker within the same audio file. We divided the trials into three groups, for annotation length of 5, 10, and 15 seconds. The results of the analysis are shown in Tab. 4. We see that with the increasing segment length, the improvement is almost negligible. Five-second annotation is satisfactory for the algorithm to find the speaker segments in the enrollment audio.

The third set of experiments studies the effect of the number

Table 1: Overall results of the individual systems and their fusion for all SITW conditions.

system	**16 kHz PLP**			**16 kHz MFCC**			**8 kHz SBN+MFCC**			**Fusion**		
condition	DCF	$\mathrm{DCF}_{\mathrm{min}}$	EER [%]	DCF	$\mathrm{DCF}_{\mathrm{min}}$	EER [%]	DCF	$\mathrm{DCF}_{\mathrm{min}}$	EER[%]	DCF	$\mathrm{DCF}_{\mathrm{min}}$	EER[%]
core-core	0.6994	0.6881	9.22	0.7151	0.7132	9.34	0.5669	0.5602	7.7210	0.5060	0.5032	5.85
core-multi	0.6996	0.6986	11.40	0.7290	0.7282	11.93	0.5768	0.5607	8.8916	0.5834	0.5650	7.34
assistcln-core	0.4843	0.4619	6.55	0.5005	0.4949	6.47	0.3867	0.3622	5.2118	0.3447	0.3405	4.71
assistcln-multi	0.4863	0.4811	7.87	0.5073	0.5022	7.67	0.3980	0.3738	6.0249	0.4298	0.3997	5.71
assist-core	0.5658	0.5581	7.24	0.5857	0.5835	7.48	0.4420	0.4273	5.4619	0.4010	0.3976	4.51
assist-multi	0.5785	0.5766	8.49	0.6056	0.6032	8.85	0.4579	0.4425	6.1030	0.4662	0.4553	5.65

Table 2: Gender-dependent results of 8 kHz system based on SBN+MFCC features for *core-core* condition.

gender	DCF	$\mathrm{DCF}_{\mathrm{min}}$	EER [%]
male	0.5195	0.5163	07.60
female	0.8600	0.8023	16.82

Table 3: Results of the 8 kHz MFCC+SBN system for the *assistclean-core* condition with different parts of enrollment audio used for i-vector extraction.

spks	part of audio	DCF	$\mathrm{DCF}_{\mathrm{min}}$	EER [%]
1	no diarization	0.4216	0.3942	6.11
	annotated part	0.6248	0.6094	8.62
	diarization	0.4138	0.3843	6.24
2	no diarization	0.4754	0.4254	6.57
	annotated part	0.5798	0.5757	7.66
	diarization	0.3865	0.3437	5.06
3+	no diarization	0.6792	0.5156	4.62
	annotated part	0.5236	0.4908	5.56
	diarization	0.4123	0.3506	3.98
all	no diarization	0.4823	0.4382	6.27
	annotated part	0.5750	0.5733	7.62
	diarization	0.3867	0.3622	5.21

of speakers in the multi-speaker conditions with respect to diarization. Note that for the test-part of the trial, we do not have the assisted annotation, i.e., we cannot use the same algorithm as in Sec. 4.1. We divided the trials from the *core-multi* into three groups based on the number of speakers in the test audio. The groups are defined for 1, 2, and 3+ speakers. In Tab. 5, we see that with increasing number of speakers in the test part of the trial, the performance degrades.

5. Conclusions

In this work, we have described our effort in SITW 2016 challenge. This challenge was interesting especially for the real acoustic environment and multi-speaker condition, where diarization was a necessary part of the systems.

We have built three i-vector systems, for 8 kHz and 16 kHz audio. We have experimented with MFCCs, PLPs, and BN features. The performance of our fusion reached $5.85\,\%$ EER in the *core-core* condition. Part of the experiments was focused on the analysis of the diarization under different conditions.

Table 4: Results of the 8 kHz MFCC+SBN system for the *assistclean-core* condition with different length of annotated part in enrollment audio.

annotation time [s]	DCF	$\mathrm{DCF}_{\mathrm{min}}$	EER [%]
5	0.3863	0.3627	5.28
10	0.3894	0.3614	5.23
15	0.3862	0.3614	5.17

Table 5: Results of the 8 kHz MFCC+SBN system for the *core-multi* condition with different number of speakers in the test audio.

spks	DCF	$\mathrm{DCF}_{\mathrm{min}}$	EER [%]
1	0.5689	0.5664	7.83
2	0.5442	0.5276	8.63
3+	0.6679	0.5983	10.38

6. References

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," vol. PP, no. 99, pp. 1 –1, 2010.

[2] S. J. D. Prince, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil, 2007.

[3] Sandro Cumani, Pietro Laface, and Oldřich Plchot, "On the use of i-vector posterior distributions in probabilistic linear discriminant analysis," *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING*, vol. 22, no. 4, pp. 846–857, 2014.

[4] Oldřich Plchot, Spyros Matsoukas, Pavel Matějka, Najim Dehak, Jeff Ma, Sandro Cumani, Ondřej Glembek, Hynek Heřmanský, Nima Mesgarani, Mohammad Mehdi Soufifar, Samuel Thomas, Bing Zhang, and Xinhui Zhou, "Developing a speaker identification system for the darpa rats project," in *Proceedings of ICASSP 2013*, Vancouver, CA, 2013.

[5] Y. Lei, L. Ferrer, M. McLaren, and N. Scheffer, "Comparative study on the use of senone-based deep neural networks for speaker recognition," *Submitted to IEEE Trans. ASLP*, 2014.

[6] Karel Veselý, Martin Karafiát, František Grézl, Miloš Janda, and Ekaterina Egorova, "The language-independent bottleneck features," in *Proceedings of IEEE*

2012 Workshop on Spoken Language Technology, 2012, pp. 336–341.

[7] 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.

[8] V. Fontaine, C. Ris, J-M. Boite, and Multitel Site Initialis, "Nonlinear Discriminant Analysis for Improved Speech Recognition," in *Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece*, 1997.

[9] František Grézl, Martin Karafiát, and Lukáš Burget, "Investigation into Bottle-neck Features for Meeting Speech Recognition," in *INTERSPEECH 2009*. 2009, pp. 2947–2950, International Speech Communication Association.

[10] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot, et al., "Promoting robustness for speaker modeling in the community: the prism evaluation set," *https://code.google.com/p/prism-set/*, 2012.

[11] "National institute of standards and technology," http://www.nist.gov/speech/tests/spk/index.htm.

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

[13] Pavel Matějka, Lukáš Burget, Petr Schwarz, and Jan Černocký, "Brno university of technology system for NIST 2005 language recognition evaluation," in *Proceedings of Odyssey 2006*, San Juan, PR, 2006.

[14] P. Kenny, "Bayesian analysis of speaker diarization with eigenvoice priors," in *CRIM*, 2008.

[15] D. Reynolds P. Kenny and F. Castaldo, "Diarization of telephone conversations using factor analysis," .

[16] Kornel Laskowski and Jens Edlund, "A Snack implementation and Tcl/Tk interface to the fundamental frequency variation spectrum algorithm," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010.

[17] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. Paliwal, Eds., New York, 1995, Elseviever.

[18] Martin Karafiát, František Grézl, Karel Veselý, Mirko Hannemann, Igor Szőke, and Jan Černocký, "BUT 2014 Babel system: Analysis of adaptation in NN based systems," in *Interspeech 2014*, 2014, pp. 3002–3006.

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

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

[21] Daniel Garcia-Romero, "Analysis of i-vector length normalization in Gaussian-PLDA speaker recognition systems," 2011.

[22] "Bosaris toolkit," https://sites.google.com/site/bosaristoolkit/.

[23] "The 2010 NIST speaker recognition evaluation plan (SRE10)," http://www.itl.nist.gov/iad/mig/tests/sre/2010/.