INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

IITG-Indigo System for NIST 2016 SRE Challenge

Nagendra Kumar[1], *Rohan Kumar Das*[1], *Sarfaraz Jelil*[1], *Dhanush B K*[2], *H. Kashyap*[2], *K. Sri Rama Murty*[3], *Sriram Ganapathy*[2], *Rohit Sinha*[1] *and S. R. M. Prasanna*[1]

[1]Indian Institute of Technology Guwahati, Guwahati-781039, India.
[2]Indian Institute of Science, Bangalore-560012, India.
[3]Indian Institute of Technology Hyderabad, Sangareddy-502285, India.

{k.nagendra, rohankd, sarfaraz, rsinha, prasanna}@iitg.ernet.in;
{dhanush.bk93, h.harishkashyap}@gmail.com; ksrm@iith.ac.in; sriram@ee.iisc.ernet.in

Abstract

This paper describes the speaker verification (SV) system submitted to the NIST 2016 speaker recognition evaluation (SRE) challenge by Indian Institute of Technology Guwahati (IITG) under the fixed training condition task. Various SV systems are developed following the idea-level collaboration with two other Indian institutions. Unlike the previous SREs, this time the focus was on developing SV system using non-target language speech data and a small amount unlabeled data from target language/dialects. For addressing these novel challenges, we tried exploring the fusion of systems created using different features, data conditioning, and classifiers. On NIST 2016 SRE evaluation data, the presented fused system resulted in actual detection cost function (*actDCF*) and equal error rate (*EER*) of 0.81 and 12.91%, respectively. Post-evaluation, we explored a recently proposed pairwise support vector machine classifier and applied adaptive S-norm to the decision scores before fusion. With these changes, the final system achieves the *actDCF* and *EER* of 0.67 and 11.63%, respectively.

Index Terms: pairwise SVM, IFCC, KDA, AS-norm.

1. Introduction

The NIST speaker recognition evaluation (SRE) is targeted towards developing a practical speaker verification (SV) system. In NIST 2016 SRE, the focus is on the use of non-target language speech in developing the SV systems. The data for system development is extracted from LDC's Call My Net (CMN) collection and comprises of speaker data in Tagalog and Cantonese (referred to as *major* language) and Cebuano and Mandarin (referred to as *minor* language). The NIST 2016 SRE, the evaluation data is taken from the major languages while the development test set is derived from the minor languages. Unlike the SRE 2012 [1], it limits the data used for developing the enroll model. For 75% of the classes only 1 utterance is provided while the remaining have 3 utterances. All enroll utterances are of approximately 60 seconds while the durations of the test utterances vary uniformly between 10 to 60 seconds. There are no cross-gender or cross-language trials in the evaluation set. The data from past SREs and other English corpora are allowed to be used in the system development. The NIST 2016 SRE plan [2] invited the participation for two tasks: fixed and open training conditions. Unlike the fixed condition, the data from a few other designated sources could be incorporated in the system development.

This paper presents the details of SV systems submitted by Indian Institute of Technology Guwahati (IITG) under the fixed training condition of the NIST 2016 SRE. These systems are developed on the basis of idea-level collaboration with two other Indian institutions: Indian Institute of Technology Hyderabad and Indian Institute of Science Bangalore. The fused system, referred to as *IITG-Indigo system*, formed the final system. The paper is organized as follows. The details of developed systems are given in Section 2. Our efforts made towards improving the system performance post-evaluation are described in Section 3. The revised system performances are given in Section 4. Finally conclusion is given in Section 5.

2. Description of the Developed Systems

This section describes SV systems developed for fixed training condition of the SRE 2016 by incorporating diversity in terms of front-end features, development data and classifier.

2.1. Speech Database

The NIST 2016 SRE evaluation data consists of 1202 enroll and 9294 test speech utterances. For evaluation, 1.9 million trials derived from the test data are used. It has been further divided into 16 parts based on the gender, the language type, the number of utterances used in creating the enroll models and the data recording setup. Unlike previous SREs, no gender information is provided for the evaluation data. For system development, the data derived from the previous SREs, the Switchboard-1 corpus, the Fisher corpus and the CMN corpus are used. Table 1 presents the details of the same.

2.2. Signal Processing and VAD

All the speech utterances used in these SV systems are sampled at a frequency of 8 kHz with 16 bits/sample resolution. The short-time analysis of speech is done using Hamming window of duration 20 ms with a frame shift of 10 ms. A likelihood-ratio test over speech and non-speech models with two thresholds has been employed to separate the speech frames from the silence frames [3]. The utterances in the development data are redistributed to have an average duration of 3 minutes after voice activity detection (VAD). The trials of minor languages are used for dev-trials. These descriptions remain common to all the developed systems. The front-end feature kind, number of Gaussian and scoring details are given in Table 2.

2.3. System-1

The first system is based on the i-vector [4] based modeling with kernel discriminant analysis (KDA) [5] as feature discrimination. It uses the instantaneous frequency cosine coefficient (IFCC) as features. The IFCC is an attempt to extract features

Copyright © 2017 ISCA

http://dx.doi.org/10.21437/Interspeech.2017-1307

Table 1: *The description of the specific corpora used in the developemnt of different SV systems.*

Notation	Corpus Name, Language	Notation	Corpus, Language
SW1	Switchboard Corpus-1, English	S08	SRE 2008, English
FRC	Fisher Corpus, English	S10	SRE 2010, English
S04	SRE 2004, English	S12	SRE 2012, English
S05	SRE 2005, English	CMN	Call My Net, Chinese
S06	SRE 2006, English		
System-1	S06+S08+S10+S12	System-3	SW1+FRC+S04+...+S12+CMN
System-2	SW1+FRC+S04+...+S12+CMN	System-4	SW1+S08

Table 2: *System details highlighting the diverse features incorporated in the component SV systems.*

System	Front-end features	UBM size	i-vector dimension	Scoring
System-1 (S1)	20IFCCs+Δ+$\Delta\Delta$	2048	600	KDA-CDS
System-2 (S2)	19MFCCs+C_0+Δ+$\Delta\Delta$	1024	500	Gaussian PLDA
System-3 (S3)	19MFCCs+C_0+Δ+$\Delta\Delta$	1024	500	Gaussian PLDA
System-4 (S4)	20MFCCs+Δ+$\Delta\Delta$	2048, 4677†	400, 300*	LDA-WCCN-PLDA

†size of DNN-UBM posterior; * dimensionality of DNN based i-vector

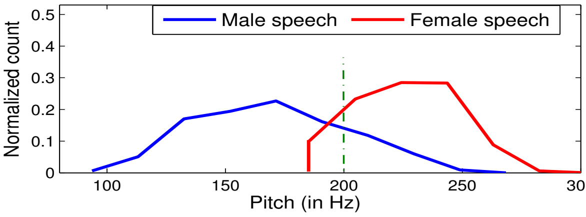

Figure 1: *The pitch value histogram of the CMN minor language labeled speech data for a total of* 1327 *utterances.*

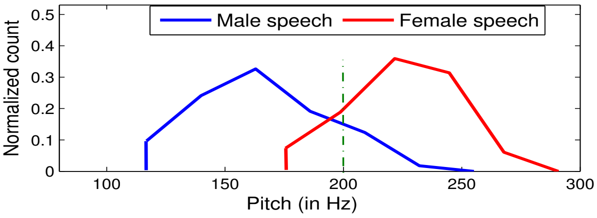

Figure 2: *The histogram of pitch value calculated for evaluation enroll data. The pitch based gender classification achieves* 86% *accuracy when keeping the threshold* 200 *Hz.*

from the analytic phase of speech signal for speaker verification [6]. In order to overcome the problem of phase warping, instantaneous frequency (IF) is computed with the help of Fourier transform properties without explicit involvement of computation of analytic phase. The narrow band components of speech are taken to compute IF by the following way,

$$\theta'[n] = \frac{2\pi}{N} Re\left\{ \frac{F_d^{-1} kZ[k]}{F_d^{-1} Z[k]} \right\} \tag{1}$$

where, F_d^{-1} denotes inverse discrete Fourier transform (DFT), N being the length of the narrow band signal and $Z[k]$ is the DFT of the analytic signal $z[n]$, obtained from the narrowband component of speech signal as explained in [7].

The computation of IF is followed by discrete cosine transform on deviations in IF computed from narrowband components of speech to extract IFCC features as a compact representation [6]. The IFCC features are found to be robust against the mismatch of vocal efforts than compared to features like mel frequency cepstral coefficients (MFCC) and frequency domain linear prediction (FDLP). Also as IFCC features contain the phase information which is not present in MFCC and FDLP features, their fusion with each of them helps in achieving improved performance as reported in [6].

2.4. System-2

This SV system employs a gender-dependent i-vector Gaussian probabilistic linear discriminant analysis (GPLDA) [8] framework. Both the unlabeled CMN and the development/evaluation enrollment data are partitioned into pseudo male and female based on the estimate of the average pitch of the utterances. For doing the same, the frame-wise pitch was extracted for 1327 utterances corresponding to gender-labeled minor language in the CMN data using the VOICEBOX suite. The average pitch was computed excluding the silence frames. The histogram of the average pitch of all gender-labeled minor languages utterances in the CMN utterance is shown in Figure 1. The enroll utterances having the pitch value less than 200 Hz are considered as belonging to pseudo-male while the remaining are treated as pseudo-female. The chosen threshold value is heuristically optimized to minimize the gender labeling error on the minor language data. From the histogram of the pitch values of the unlabeled major languages in the enroll data shown in Figure 2, it is evident that the chosen threshold is effective for this case too. Post-evaluation, using the released key, we found that the developed scheme achieved an gender classification accuracy of 86% for the unlabeled enroll data.

For each partition of the unlabeled CMN development data, a 1024 Gaussian UBM is created. Based on the estimated gender information of enrollment data, $10,393$ dev-trials and $657,109$ eval-trials are selected as male-speaker trials, remaining are treated as female-speaker trials. The initial system for tuning the parameters is learned by excluding the CMN development data. A total of $61,000$ utterances amounting to 3100 hours of telephone recorded speech data after VAD is available for learning the total-variability (TV) matrix. For gender-based modeling, separate TV-matrices are created using $25,000$ male and $36,000$ female utterances. In GPLDA modeling, the i-vectors, applied with whitening and length normalization, are mapped to 400-dimensional subspace. Separate GPLDA systems are developed for each broad partitioning of the data but using previous SREs data only.

2.5. System-3

As mentioned earlier, all development data utterances are in English language except for very small amount of Chinese language data taken from the CMN dataset. In SRE 2016, the enroll and test data comes from Chinese language dialects. To minimize this difference, we attempted to map all the English spoken training data MFCC vectors are mapped using the unlabeled data. For the same a gender-dependent K-mean singular value decomposition [9] dictionary $\boldsymbol{D}$ of 2000 atoms is learned using pooled unlabeled major and minor language data. The MFCC feature vectors $\mathbf{y}$ of all English development data are sparse coded over the created dictionary $\boldsymbol{D}$ as

$$\hat{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}} \|\boldsymbol{y} \; - \; \boldsymbol{D}\boldsymbol{u}\|_2 \quad \text{subject to} \quad \|\boldsymbol{u}\|_0 \leq l \qquad (2)$$

where $\boldsymbol{u}$ is the vector of unknown coefficients and l is the chosen sparsity constraint. For finding the sparse solution $\hat{\boldsymbol{u}}$, the orthogonal mapping pursuit [10] algorithm with the sparsity value of 10, is used. Using so obtained sparse vector, the target vector is re-synthesized as $\tilde{\boldsymbol{y}} = \boldsymbol{D}\hat{\boldsymbol{u}}$. The synthesized MFCC feature vectors are nothing but sparse linear combination of the dictionary atoms derived from Chinese language data. As a result of that, these vectors are expected to reduce the acoustic mismatch between development and testing data and are referred to as *mapped* MFCC feature vectors in this work. An alternate SV system, designated as System-3, is developed using the *mapped* MFCC features in learning the TV-matrices while keeping the rest of attributes identical to earlier discussed System-2.

2.6. System-4

This system consists of fusion of two subsystems. The first subsystem is similar to System-1 with a 2048-component UBM and 400 dimensional i-vectors. Dimensionality reduction is performed using linear discriminant analysis (LDA) and within-class covariance normalization (WCCN) [11] to reduce the i-vector dimensions to lower subspace. Subsequently, PLDA [12] is used to score the evaluation trials. The second subsystem is built using deep neural network (DNN) based posteriors [13,14]. The posteriors are used to create a 4677-component UBM and 300 dimensional i-vectors are computed. Dimensionality reduction is performed using LDA-WCCN and scoring is done using PLDA. The scores of the aforementioned subsystems are fused. NIST SRE 2008 and switchboard corpora are used in the development of both the subsystems.

2.7. System Performance

The performance is primarily evaluated in terms of actual detection cost function (actDCF) as per NIST 2016 SRE protocol [2], and corresponding minimum DCF (minDCF) and equal error rate (EER) are given for comparison purpose. The actDCF reflects both discriminative and calibration ability of the SV system, while the minDCF shows only discriminative ability of the SV systems. The probability of target is kept 0.01 and 0.005 for computing the actDCF. The remaining parameters for computing the actDCF is fixed such that, it computes the performance in the low false alarm region. The performances of the different SV systems and their fusion are computed using the scoring script released by the NIST. The scores calibration and fusion parameters are learned using the BOSARIS toolkit [15].

Table 3: *SV performances of the component systems and their fusion that was presented at NIST 2016 SRE Workshop.*

System	actDCF	% EER
System-1: S1	0.962	21.55
System-2: S2	0.922	14.44
System-3: S3	0.907	12.60
System-4: S4	0.915	16.11
Fusion: all sys	0.806	12.91

Table 3 shows the performance of component systems and their fusion on SRE 2016 evaluation data. The significant gain achieved on fusion is attributed to diversity of the component systems. Further, it would be interesting to find how effective was the global feature mapping explored for addressing the mismatch between the languages involved in the evaluation and the system development data. On comparing the performances of S2 and S3 system, we note that feature mapping has resulted in a relative improvement of 1.63% and 12.74% in actDCF and EER, respectively.

3. Post Evaluation Work

Inspired the systems presented in SRE 2016 Workshop, we explored some new features in our developed system. In particularly, we have explored the adaptive S-norm (AS-norm) [16] for the score normalization and built a new component SV system employing the recently proposed pairwise support vector machine (PSVM) [17]. In the following, we first briefly describe AS-norm and PSVM; following that their impact of the system performance is presented.

3.1. Adaptive S-Norm

In pattern recognition problems, the score normalization is usually employed to transform the raw scores such that a common threshold can be effectively used for classification. The AS-norm is derived from the adaptive T-norm [18] but it also preserves the symmetrical property of the S-norm [19]. The AS-norm score s' for a verification score s obtained from two speech utterances i_1 and i_2 is computed as following:

$$s' = \frac{1}{2}\Big[\frac{s-\mu_1^{\boldsymbol{n}_2}}{\sigma_1^{\boldsymbol{n}_2}} + \frac{s-\mu_2^{\boldsymbol{n}_1}}{\sigma_2^{\boldsymbol{n}_1}}\Big] \qquad (3)$$

where $\mu_1^{\boldsymbol{n}_2}$ and $\sigma_1^{\boldsymbol{n}_2}$ are the mean and standard deviation computed using the normalization imposter subset $\boldsymbol{n}_2$ depends on the speech utterance i_2, and the same notation duality is applied to the second term. In this work, the normalization imposter subsets ($\boldsymbol{n}_1$ and $\boldsymbol{n}_2$) are derived selected based on top absolute correlated the claimed speaker utterances (i_1 or i_2) and the imposter utterances. The all unlabeled data from minor and major language is selected as imposter utterances for score normalization. The number of imposter utterances selected by AS-norm for computing the mean and variance is set to 250.

3.2. Pairwise Support Vector Machine

There are two ways to do multiclass classification using the support vector machine (SVM) termed as "one-vs-one" and "one-vs-rest". These approaches learn separate model for each class enroll data and it requires sufficiently large number of enroll data per class to achieve robust classification performance. Recently, a new discriminative SVM model named as pairwise

Table 4: *Performances of different SV systems developed along with fusions of the systems under the fixed condition task on the evaluation test set of NIST 2016 SRE. System-2, System-3 and System-5 use the AS-norm.*

SV System	**actDCF**					**minDCF**	**%EER**
	Cantonese		Tagalog		**Avg.**	**Avg.**	**Avg.**
	1-utt	3-utt	1-utt	3-utt			
S1	0.942	0.764	1.083	1.059	**0.962**	0.939	21.55
S2*	0.780	0.611	0.963	0.903	**0.818**	0.815	14.25
S3*	0.845	0.662	0.967	0.953	**0.857**	0.810	12.52
S4	0.873	0.892	0.951	0.944	**0.914**	0.904	16.11
S5*	0.766	0.556	0.943	0.873	**0.784**	0.783	14.58
F1: S2*+S3*	0.711	0.518	0.931	0.858	**0.754**	0.737	12.51
F2: S1+S3*+S4	0.646	0.450	0.956	0.811	**0.716**	0.712	12.35
F3: S2*+S3*+S4+S5*	0.612	0.416	0.887	0.775	**0.672**	0.670	11.63

* Indicates that the system score is normalized with AS-norm.

SVM (PSVM) [17] is proposed where a single model is used for multiclass classification. The PSVM algorithm converts the multiclass classification problem to 2-class. The PSVM approach relies on two input and it predicts whether pair inputs belong to the same class or not. Its performance remains robust even if training data exclude the enroll data. Learning of the PSVM model does not require the class label of the training data but it need pair information, whether they belong to the same class or not. In other words, it learns the two class models as "same-class" vs "different-class" using all the training pairs. In this way, the same class training pairs grow linearly with number of class while different class training pairs grows quadratically with the number of utterances. For addressing that, an efficient version of the PSVM algorithm is proposed in [20, 21] that discards non-contributing support vectors. The objective of PSVM in its primal form can be defined as risk minimization problem

$$F(\boldsymbol{\alpha}) = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2}\lambda\|\boldsymbol{\alpha}\|^2 + \frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{\alpha}, \boldsymbol{x}_i, \eta) \qquad (4)$$

where n is the number of training pairs used for learning the PSVM model, $\boldsymbol{x}_i$ is the paired training data associated with label whether they belong to same class or not and η is regularization factor. There are two ways to learn the PSVM model either symmetric kernel can be used or the symmetric training pair can be used with no constraint on the kernel. If we have two data utterances i_1 and i_2, then the symmetric training pair approach uses both pairs (i_1, i_2) and (i_2, i_1), where as symmetric kernel uses any one pair. It is shown in [17], both types of training approaches lead to the same performance.

3.3. System-5

This system is developed post evaluation and incorporates symmetric linear kernel based PSVM [20, 21] approach for classification. The rest of details of this system is identical to those of System-2. The gender-wise PSVM model is trained using the i-vectors corresponding to the data taken from previous SREs. For learning the PSVM model, the number of mismatched-speaker pairs is kept 50 times more than the number of matched-speaker pairs. As a result of that there are 4 million male data and 6.1 million female data training pairs are generated. In [20, 21], the authors employed the PLDA likelihood scores in training pair formation. Unlike that, we have used the correlation scores in training pair formation. This reduces the computational complexity of the pair-formation substantially without any significant loss in the performance.

4. Refined System Performances

Before discussing about the post-evaluation efforts on the system performances, we wish to highlight the robustness of developed SV systems by giving the breakup of the performances in term of salient factors present in the evaluation data. Table 4 shows the SV performances divided into four parts based on the dialect (Tagalog or Cantonese) and the number of utterances (1 or 3) used in creating the enroll model. As expected, the performances for 3 utterance enroll model turn out to be better than those for single utterance enroll model. Further, on comparing the performances across dialects, Cantonese is noted to be easier than Tagalog. Needless to mention, these comparisons are loose as the trials across these breakups are not identical.

In Table 4, the performances for S1 and S4 systems are without AS-norm. This is due to loss of the i-vectors for these cases due to a recent disk crash. The remaining component systems where the AS-norm is applied, are differentiated by appending '$*$' with their codes. On comparing the performances of S2 and S3 systems across Table 3 and Table 4, we note that a significant improvement in actDCF is noted with AS-norm. The S5 system has resulted in consistent improvement in actDCF for all the cases over the S2 system and this highlights the impact of the explored PSVM approach along with the AS-norm. The trends for the feature mapping noted earlier do not seem to hold when AS-norm is applied. This could be due to improper calibration. Instead, we can assess on the basis of minDCF and EER measures which are unaffected by the calibration error. On both these measures, the SV system with features mapping outperforms the one without it. For exploiting the diversity among the component systems various combinations of system fusion have been explored. With best possible combination an actDCF of 0.670 and EER of 11.63% is achieved.

5. Conclusion

This work presents the system description of the submission made by IITG to the NIST 2016 SRE. For meeting the challenge of lack of target language data and the limited enroll data, we explored a fusion of SV systems developed employing different feature, data conditioning and classifiers. With post-evaluation refinements, the final system is noted to yield an actual detection cost of 0.67 on SRE 2016 evaluation data.

6. Acknowledgements

The authors wish to thank Andreas Fischer for sharing the pairwise SVM code used in this work.

7. References

[1] *The NIST Year 2012 Speaker Recognition Evaluation Plan, www.nist.gov/document-6865.*

[2] *The NIST 2016 Speaker Recognition Evaluation Plan, www.nist.gov/document/sre16evalplanv1-0pdf.*

[3] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7229–7233.

[4] 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, 2011.

[5] D. Cai, X. He, and J. Han, "Efficient kernel discriminant analysis via spectral regression," in *IEEE International Conference on Data Mining, ICDM 2007*. IEEE, 2007, pp. 427–432.

[6] K. Vijayan, P. R. Reddy, and K. S. R. Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Communication*, vol. 81, pp. 54–71, 2016.

[7] L. Marple, "Computing the discrete-time analytic signal via FFT," *IEEE Transactions on Signal Processing*, vol. 47, no. 9, pp. 2600–2603, 1999.

[8] 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.

[9] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[10] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit," Technion, TR-CS-2008-08, Tech. Rep., 2008.

[11] A. O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, Sep 2006, pp. 1471–1474.

[12] S. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, Oct 2007, pp. 1–8.

[13] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1695–1699.

[14] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos, "The ibm 2016 speaker recognition system," *arXiv preprint arXiv:1602.07291*, 2016.

[15] The BOSARIS toolkit, accessed on 10th Dec. 2013. [Online]. Available: www.sites.google.com/site/bosaristoolkit/

[16] S. Cumani, P. D. Batzu, D. Colibro, C. Vair, P. Laface, and V. Vasilakakis, "Comparison of speaker recognition approaches for real applications." ISCA, 2011.

[17] C. Brunner, A. Fischer, K. Luig, and T. Thies, "Pairwise support vector machines and their application to large scale problems," *Journal of Machine Learning Research*, vol. 13, no. Aug, pp. 2279–2292, 2012.

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

[19] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Odyssey*, 2010, p. 14.

[20] S. Cumani and P. Laface, "Training pairwise support vector machines with large scale datasets," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014*. IEEE, 2014, pp. 1645–1649.

[21] S. Cumani and P. Laface, "Large-scale training of pairwise support vector machines for speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 11, pp. 1590–1600, 2014.