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

Nuance - Politecnico di Torino's 2016 NIST Speaker Recognition Evaluation System

Daniele Colibro[1], *Claudio Vair*[1], *Emanuele Dalmasso*[1], *Kevin Farrell*[1], *Gennady Karvitsky*[1], *Sandro Cumani*[2], *Pietro Laface*[2]

[1] Nuance Communications, Inc.
[2]Politecnico di Torino, Italy

{Daniele.Colibro,Claudio.Vair,Kevin.Farrell,Emanuele.Dalmasso,Gennady.Karvitsky}@nuance.com
{Sandro.Cumani,Pietro.Laface}@polito.it

Abstract

This paper describes the Nuance–Politecnico di Torino (NPT) speaker recognition system submitted to the NIST SRE16 evaluation campaign. Included are the results of post-evaluation tests, focusing on the analysis of the performance of generative and discriminative classifiers, and of score normalization. The submitted system combines the results of four GMM-IVector models, two DNN-IVector models and a GMM-SVM acoustic system. Each system exploits acoustic front-end parameters that differ by feature type and dimension. We analyze the main components of our submission, which contributed to obtaining 8.1% EER and 0.532 actual $C_{primary}$ in the challenging SRE16 Fixed condition.

Index Terms: Speaker Recognition, i-vector, PLDA, PSVM, AS-Norm, Top-Norm.

1. Introduction

The 2016 Speaker Recognition Evaluation (SRE16) organized by the National Institute of Standards and Technology (NIST), focuses on the speaker detection task. As usual, the goal is to decide whether a target speaker is speaking in a segment of conversational speech. The main difference of the 2016 evaluation with respect to prior evaluations is that the enrollment and test segments were recorded outside North America, in Tagalog and Cantonese. The new corpus collection also includes a development set, which consists of an unlabeled component of the two evaluation languages (referred to as Major languages by NIST), plus 10 labeled segments collected from 20 speakers of two additional oriental languages, Cebuano and Mandarin (referred to as Minor languages). Furthermore, NIST defined a mandatory "Fixed" training condition that constrains the data used for developing the evaluation systems to come exclusively from previous SREs, Switchboard and Fisher corpora and from the development component of the new collection.

SRE16 includes two enrollment conditions, using one or three segments, respectively. These segments contain a nominal speech amount of 60 seconds, whereas the test segments have duration uniformly sampled from 10 to 60 seconds.

NIST defined the 2016 Detection Cost Function (DCF) as the combination of two costs having different target probabilities and additionally where the detection costs across different conditions were equalized. The conditions for equalization include one and three-segment enrollment, Tagalog and Cantonese, male and female, and same and different phone number. A detailed description of the data, tasks, rules, and DCF of SRE16 can be found in the evaluation plan available in [1].

This paper presents the strategic choices, and describes the techniques that we found useful for obtaining a successful evaluation system. In particular, we highlight the most relevant factors that contributed to accuracy improvements within our speaker recognition systems. Furthermore, we analyze the performance of generative and discriminative classifiers, together with the impact of different approaches to score normalization.

The paper is organized as follows: Sections 2 illustrates the architecture of the system, the voice activity detection and feature extraction modules, and the speaker models. Section 3 describes our development setup, the adaptation strategy, the scoring technique, and the results obtained on the development data. Experimental results on the evaluation set, and post-evaluation considerations are given in Section 4. Conclusions are drawn in Section 5.

2. System architecture

The main modules and techniques that have been designed and tuned for this evaluation include Voice Activity Detection (VAD), feature extraction, feature normalization, and speaker recognition. These components are described, in turn, in the following subsections.

2.1. Voice activity detection

We used two VAD models based on Neural Network (NN) phonetic decoding. The decoders are hybrid HMM-NN models trained to recognize 11 phone classes [2] or detailed English-US acoustic units.

The Neural Networks used for the VAD are Multilayer Perceptrons that estimate the posterior probability of phonetic units (or classes), given an acoustic feature vector. They have been trained on Switchboard and Fisher corpora using approximately 200 hours of speech.

2.2. Feature extraction and normalization

Six sets of features have been extracted for training the models used in this evaluation. Some of these subsystems exploit feature warping by means of short term gaussianization [3]. The other systems perform cepstral mean and variance normalization on the entire audio segment. Table 1 summarizes the set of acoustic features that have been used.

http://dx.doi.org/10.21437/Interspeech.2017-797

Table 1. *Acoustic features and normalization techniques: Gaussianization (GAU) or Cepstral Mean and Variance Normalization (CMVN)*

Feature type	Feature number	Features	Norm
MFCC60	60	20 MFCC + 20 Δ + 20 ΔΔ	GAU
PLP60	60	20 PLP + 20 Δ + 20 ΔΔ	GAU
PLPCMN60	60	20 PLP + 20 Δ + 20 ΔΔ	CMVN
MFCC&PLP	60	15 MFCC + 15 Δ 15 PLP + 15 Δ	GAU
MFCC45	45	18 MFCC (no c0) + 19 Δ + 8 ΔΔ	GAU
PLPCMN45	45	19 PLP +19 Δ + 7 ΔΔ	CMVN

Analysis bandwidth, window lengths, number of Mel filters, liftering, etc., have been configured with different values for the considered feature extractors, in order to maximize the feature orthogonality.

Short time gaussianization and cepstral mean and variance normalization were performed for each static parameter, on the frames selected by the VAD module.

2.3. Speaker recognition

We used for our submission the combination of three different models, two classifiers, and up to four different acoustic features:

- GMM-IVector with Pairwise SVM (4 systems)
- DNN-IVector with Pairwise SVM (2 systems)
- GMM-SVM with NAP (1 system)

2.3.1. GMM-IVector with Pairwise SVM

The GMM-IVector extractors follow the standard paradigm proposed in [4].

We trained gender independent UBMs with 2048 diagonal covariance matrices, and total variability T matrices with 500 factors by means of Expectation-Maximization iterations, using the telephone segments of SRE04-SRE10 and Switchboard.

The speaker recognition raw scores have been obtained by using a Pairwise Support Vector Machine (PSVM) [5], [6], trained on NIST SRE data. For some systems, the SRE segments have been cut in order to limit the speech duration between 10 and 60 seconds, obtaining multiple segments per audio file. PSVM training has been performed using approximately 100K training segments.

2.3.2. DNN-IVector with Pairwise SVM

The DNN-IVector extractors are based on the hybrid Deep Neural Network/GMM approach for extracting Baum-Welch statistics proposed in [7].

We used two different DNN systems trained with SRE and Switchboard data collections. The first one uses a DNN system with 6144 output units trained on Switchboard and Fisher datasets. The DNN input consists of a sliding window of 13 frames. Each frame includes 19 MFCC [8] and 12 RASTA-PLP [9] parameters. The DNN has four hidden layers with 2048 rectified linear units, and a pre-final layer with 512 nodes. The final Softmax layer produces the posterior probability of the output units. Based on the DNN posterior probabilities, Baum-Welch statistics for training the Gaussian UBM and the T matrix were computed on 45 MFCC features.

The second DNN-IVector model uses the same acoustic features as the previous system, but it is based on a DNN with less parameters and output units: 1024 output units and 4 hidden layers with 384, 384, 384, and 128 nodes, respectively. For this DNN, the activation function for all the hidden layers is the Sigmoid. The GMM-UBM of this system relies on 60 PLP acoustic features, rather than on the 45 MFCC features.

It is worth noting that to introduce a further element of complementarity among systems, the first DNN system produce e-vectors [10], rather than i-vectors. Also in this case, trial vector pairs are scored by PSVM classifiers.

2.3.3. GMM-SVM with NAP

Our submission also includes a GMM-SVM system trained on the unlabeled components of the CallMyNet [11] development set. This system uses Nuisance Attribute Projection (NAP) [12] for mitigating the effect of intra-speaker variability. The system is based on 45 PLP features subject to cepstral mean and variance normalization. Due to the scarcity of training data, it uses a 512 Gaussian GMM-UBM.

The speaker labels needed for training the NAP model were obtained by means of speaker clustering on the unlabeled CallMyNet development data. Our speaker clustering procedure detected 271 clusters with 2 audio segments, 46 with 3 audio segments, 6 with 4 segments, and 1 cluster with 6 segments. We assumed that the remaining 1762 segments came from different speakers. Post evaluation tests allowed us to assess that cluster and speaker impurity [13] were 2.7% and 24.2%, respectively. The benefit of NAP was assessed on the labeled development data. We obtained approximately 8% of MinDCF16 relative reduction with respect to an equivalent model without NAP.

The GMM-UBM has been trained with the complete set of 2472 segments from Minor and Major unlabeled development data. From this set we randomly selected 1000 segments as the negative samples for the SVM training.

3. Development setup and results

This section illustrates how we exploited the development data. It provides details for the techniques that were found useful for adapting a system, trained on English only, to new languages. Additionally, we highlight the importance of score normalization for obtaining the best results.

3.1. Adaptation

We tested several unsupervised compensation and adaptation techniques for reducing the mismatch of the acoustic and the i-vector space with respect to the CallMyNet environment.

The most successful technique, which we exploited for the evaluation, simply relies on using CallMyNet development data for MAP adaptation on the GMM model of the English trained systems. This method provided a few percent of MinDCF16 relative error reduction across our set of systems. The improvement is not high, but it is worth considering that it was obtained in mismatched conditions, because the unlabeled data used for adaptation came mostly from Major languages, whereas the development test set included segments of the Minor languages only. Thus, we expected similar or better gain on the evaluation data.

Table 2. *Speaker recognition subsystems*

System	Features	VAD	Technology
S1	**MFCC60**	Pho classes	GMM-IVEC, 2048G
S2	**PLP60**	Pho classes	GMM-IVEC, 2048G
S3	**PLPCMN60**	En-Us units	GMM-IVEC, 2048G
S4	**MFCC&PLP**	En-Us units	GMM-IVEC, 2048G
S5	**MFCC45**	Pho classes	DNN-IVEC, 6144G
S6	**PLP60**	Pho classes	DNN-IVEC, 1024G
S7	**PLPCMN45**	Pho classes	GMM-SVM, 512G

Moreover, from previous experience with other languages, we know that additional training data, even of mismatched languages, are normally beneficial to our classifiers. Since there is a mismatch between the English data and the Oriental languages datasets provided for development and evaluation, we decided to train the classifiers for the evaluation adding the labeled CallMyNet development component to the training set. Since these data are scarce, they would have a negligible impact when added to the large set of English training segments. Thus, we included in the training set replicate copies of the labeled Minor languages segments, limiting their contribution to ~10% of the complete training set. This proportion seemed reasonable to account for the specificity of the Oriental languages and channels, without affecting too much the speaker discrimination capability of the original (robust) English models.

3.2. Score normalization

As in the SRE 2012 evaluation, the trial results of all our systems were subject to score normalization. In particular, the GMM and DNN i-vector systems use Adaptive Symmetric Normalization (AS-Norm) [14], whereas the GMM-SVM relies on ZT-Norm.

An additional normalization step has also been performed on the AS/ZT normalized scores, exploiting the Top-Norm approach proposed in [15].

3.2.1. Adaptive AS-Norm

We applied AS-Norm for normalizing the scores produced by each GMM or DNN i-vector classifier. S-Norm [16] computes the average between the Z-normed and T-Normed scores, considering, in turn, the trial i-vectors either as a model or as a test. AS-Norm is similar, but it is adaptive because the mean and variance parameters are computed only on a subset of k-nearest impostors in the normalization set. The k-nearest impostors are chosen by means of a vector distance comparison (city-block distance) as in [17].

Thus, a raw PSVM score comparing two i-vectors is normalized as the average of its Adaptive T-norm and Adaptive Z-norm values as:

$$s' = \frac{1}{2} \cdot \left[\frac{s - \mu_Z(N_2)}{\sigma_Z(N_2)} + \frac{s - \mu_T(N_1)}{\sigma_T(N_1)} \right]$$

where μ_Z and σ_Z are the mean and standard deviation of the scores obtained by scoring i-vector i_1 against the subset N_2 of the k-nearest impostors of i-vector i_2, and μ_T and σ_T are computed by analogy, inverting the role of i_1 and i_2.

Our normalization set for SRE16 consists of all the 2472 segments available in the CallMyNet unlabeled development set, including Major and Minor languages, and k, the number of selected elements was set to 200. AS-Norm provided up to 25% relative MinDCF16 reduction on our development data.

Table 3. *Unbiased scoring on the development set*

Scoring type	%EER	Min $C_{primary}$	Act $C_{primary}$
Equalized	15.9	0.568	0.582
Unequalized	15.5	0.538	0.553

3.2.2. ZT-Norm

The scores of the GMM-SVM subsystem were normalized by means of ZT-Norm (Z-Norm followed by T-Norm), using the CallMyNet unlabeled development set. We used unsupervised speaker labels, obtained through automatic speaker clustering, for randomly selecting one segment per speaker.
The normalization set includes 1000 segments. When applying ZT-Norm on the GMM-SVM, we estimated a relative improvement of approximately 5% for the MinDCF16.

Top-Norm

Tests performed on CallMyNet development data showed that additional improvement in the DCF region of interest could be achieved by exploiting the Top-Norm approach on all systems. We applied Top-Norm selection along with mean and variance normalization on the AS/ZT normalized scores. A set of the 200 best scoring impostors were selected for each pair of enrollment/test segments. Top-Norm provided, on the development data, an additional MinDCF16 relative reduction in the 2-4% range.

3.3. Score combination and calibration

For this evaluation, we used the seven systems and features, described in the previous sections, and summarized in Table 2.

The combination of these systems was obtained by linear fusion of their scores. The fusion weights and the calibration offset were computed by means of the Focal toolkit [18]. We computed different weights and calibration parameters for the two enrollment conditions of this evaluation, namely "one-segment" and "three-segment". For fusion and calibration purposes, we exploited the labeled development data, by creating new trials that were not included in the official development list. For example, we considered the enrollment segments of the "three-segment" condition for creating 3 additional "one-segment" models.

For the "one-segment" enrollment condition, we ran the Focal toolkit on a corresponding condition of the development data, which included 10362 true speaker trials and 41448 impostor trials.

The weights of the "three-segment" condition were, instead, computed by merging the "one-segment" and "three-segment" development trials because the number of development trials for the "three-segment" condition alone was insufficient. However, the calibration offset was estimated on the matching "three-segment" development set, keeping the fusion weights obtained from pooled data. It is worth noting that for avoiding the bias due to training/calibration data overlap, the calibration weights have been computed on the labeled development data using a version of the PSVM models trained without the CallMyNet dataset.

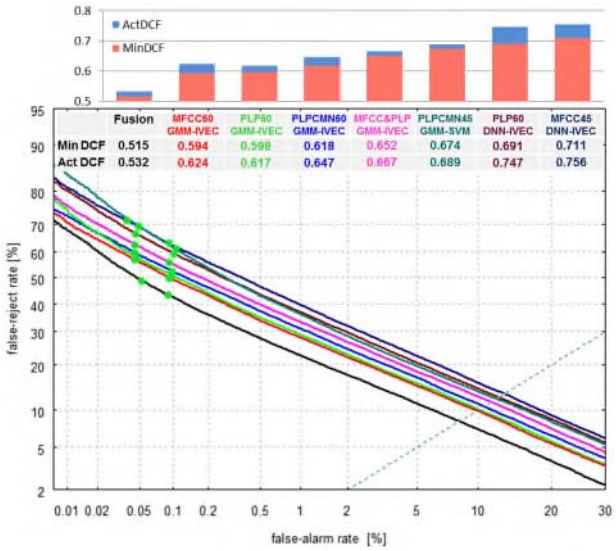

Figure 1: *Minimum and actual DCF16 and DET plots: individual contribution and fusion of seven systems.*

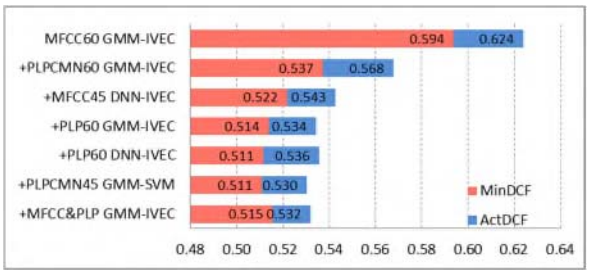

Figure 2: *Contribution of additional systems to fusion.*

Table 3 summarizes the results of the systems used for calibrating our submission on the official SRE16 development test-set. These results, obtained by using the Python scoring tool provided by NIST, are unbiased as they do not include labeled development data in the PSVM and NAP training sets.

4. Results on the evaluation set

In this section, we summarize the results obtained on the evaluation. We analyze the contribution of the different techniques and systems that have been combined, and we share some considerations about the experience that we have gained participating in this year's evaluation.

The PSVM classifier was preferred to a PLDA classifier because, in development, the former has shown to provide better accuracy.

Figure 1 shows the performance of the individual systems in terms of minimum and actual DCF16, and their DET plots. It can be noticed that the two hybrid DNN/GMM systems do not perform as well as the purely acoustic systems. Even the GMM-SVM does 7.8% better than the best DNN system on the actual DCF16. This is not surprising considering that the DNN was trained on English data only.

Looking also at Figure 2, it can be noticed that the first and third GMM-IVector systems, which use different features and normalization techniques, provide most of the improvement (9.6% of the min DCF16). The additional combination with the other most complementary system (MFCC45 DNN-IVector), allows gaining another 3%.

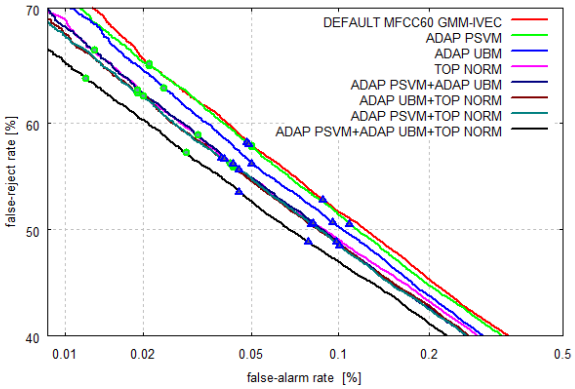

Figure 3: *DET plot comparing the contribution of UBM and PSVM adaptation, and of Top-Norm.*

Table 4. *Minimum DCF16 for three systems as a function of the scoring normalization technique.*

Normalization	MFCC60 GMM-IVEC	PLPCMN60 GMM-IVEC	MFCC45 DNN-IVEC	Fusion
No	0.949	0.967	1.000	0.910
Top-norm	0.632	0.668	0.740	0.569
AS-norm	0.629	0.651	0.743	0.551
AS+Top norm	0.594	0.618	0.711	0.522

The remaining systems give further, but smaller, contribution to the overall performance improvement.

The DET plots of Figure 3 illustrate the contribution of UBM and PSVM adaptation and of Top-Norm scoring with respect to the best individual system (MFCC60 GMM-IVector). While the adaptation of the PSVM alone does not produce any appreciable performance improvement, the UBM adaptation is essential for reducing the acoustic mismatch between the English and the evaluation environments. It is interesting that Top-Norm alone gives approximately the same contribution of the two other combined adaptations, but that using all of them yields another notable improvement as observed in the DET plot.

A comparison of the minimum DCF16 obtained by the best combination of three systems, as a function of the normalization technique that has been used, is shown in Table 4. It can be observed that AS-Norm and Top-Norm alone give a similar and relevant relative performance improvement (37.5%) with respect to raw scoring. The relative improvement increases to 42.6% by applying in sequence these two normalizations.

5. Conclusions

We presented the components and analyzed the results of the Nuance–Politecnico di Torino (NPT) speaker recognition system submitted to the NIST SRE16 evaluation campaign. We have shown that the keys for the success of our submission were 1) the use of systems based on complementary features and technology, 2) careful adaptation of the models to the Oriental languages, and 3) score normalization plus calibration dependent on the number of enrolment segments. This enabled a series of incremental improvements which yielded a substantial reduction to the error rate of our best single system.

6. References

[1] Speaker Recognition Evaluation 2016 at National Institute of Standards and Technology, Multimodal Information Group, https://www.nist.gov/itl/iad/mig/speaker-recognition-evaluation-2016

[2] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, C. Vair, "Compensation of Nuisance Factors for Speaker and Language Recognition", IEEE Trans. on Audio, Speech, and Language Processing. Vol. 15-7, pp. 1969-1978, 2007.

[3] J. Pelecanos, and S. Sridharan, "Feature Warping for Robust Speaker Verification", in Proc. 2001: A Speaker Odyssey, pp. 213-218, 2001.

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification", in IEEE Transactions on Audio, Speech, and Language Processing, Vol.19, n. 4, pp. 788-798, 2011.

[5] S. Cumani, P. Laface, "Training pairwise Support Vector Machines with large scale datasets", in 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP 2014, Florence (Italy), pp. 1664-1668, 2014.

[6] S. Cumani, P. Laface, "Large scale training of Pairwise Support Vector Machines for speaker recognition", in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22 No. 11, pp. 1590-1600, 2014.

[7] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network", in Proc. ICASSP 2014, pp. 1714–1718, 2014.

[8] S. B. Davis and P. Mermelstein: "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustic, Speech, and Signal Processing, Vol. 28, No. 4, pp. 357–366, 1980.

[9] H. Hermansky, N. Morgan ''Rasta Processing of Speech'', IEEE Trans. On Speech and Audio Proc. Vol.2, No.4, pp. 578-589 1994.

[10] S. Cumani, and P. Laface, "E–vectors: JFA and i–vectors revisited", in Proc. ICASSP 2017, New Orleans, USA, pp. 5435-5439, 2017.

[11] K. Jones, S. Strassel, K. Walker, D. Graff, J. Wright, "Call My Net Corpus: A Multilingual Corpus for Evaluation of Speaker Recognition Technology", in Proc. Interspeech 2017.

[12] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation", in Proc. ICASSP 2006, pp. 97–100, 2006.

[13] J. Jorrín Prieto, C. Vaquero, P. García, "Analysis of the Impact of the Audio Database Characteristics in the Accuracy of a Speaker Clustering System", in Odyssey 2016 The Speaker and Language Recognition Workshop, Bilbao, pp. 393-399, 2016.

[14] S. Cumani, P.D. Batzu, D. Colibro, C. Vair, P. Laface, V. Vasilakakis, "Comparison of Speaker Recognition Approaches for Real Applications", Interspeech 2011, Florence, Italy, pp. 2365-2368, 2011.

[15] Y. Zigel and M. Wasserblat, "How to Deal with Multiple Targets in Speaker Identification systems?", in Odyssey 2006 The Speaker and Language Recognition Workshop, San Juan, 2006

[16] P. Kenny, "Bayesian speaker verification with heavy tailed priors," Keynote presentation, Odyssey 2010, The Speaker and Language Recognition Workshop, 2010.

[17] D. E. Sturim, D. A. Reynolds, "Speaker Adaptive Cohort Selection for T-norm in Text-Independent Speaker Verification", in Proc. ICASSP 2005, pp. 741-744 , 2005

[18] Available at https://sites.google.com/site/nikobrummer/focal