Nuance - Politecnico di Torino's 2012 NIST Speaker Recognition Evaluation System

Daniele Colibro[1]*, Claudio Vair*[1]*, Kevin Farrell*[1]*, Nir Krause*[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,Nir.Krause,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 SRE12 evaluation campaign. Included are the results of post-evaluation tests, focusing on the analysis of the effects of score normalization and condition-dependent calibration. The submitted system combines the results of five acoustic recognizers all based on Gaussian Mixture Models (GMMs). Each system has its own front end, with features differing by their type and dimension. We illustrate the process of development data selection and configuration of state-of-the art technology, which contributed to obtaining good performance in all the test conditions proposed in this evaluation.

Index Terms: Speaker Recognition, i-vectors, Probabilistic Linear Discriminant Analysis, AS-Norm

1. Introduction

The 2012 Speaker Recognition Evaluation (SRE12) organized by the National Institute of Standards and Technology (NIST), focuses on the speaker detection task. The goal is to decide whether a target speaker is speaking in a segment of conversational speech. The main difference of the 2012 evaluation with respect to the previous ones is that most target speakers were taken from all previous SRE corpora. Furthermore, some of the test segments had additive noise imposed, and knowledge of all targets was allowed in computing each trial's detection score. System performance has been assessed using a new Detection Cost Function (DCF) defined in the evaluation plan [1] as the combination of two costs, one using the cost parameters from SRE10 and one using a larger target prior.

SRE12 included 3 training and 5 testing conditions, but only 9 different test configurations. One of these was the core, or mandatory, condition and included a set of excerpts from a telephone conversation or interview containing nominally between 20 and 160 seconds of target speaker speech. A detailed description of the data, tasks and rules of SRE12 can be found in the evaluation plan available in [1].

In this paper we present the techniques exploited for this evaluation and we highlight the factors most relevant to the training of good speaker models. Furthermore, we analyze the effects on the system performance of the selected normalization and calibration techniques.

The paper is organized as follows: Sections 2 illustrates the system architecture, the voice activity detection, the feature extraction and the speaker models. Section 3 describes the design of the development data. Section 4 is devoted to the classification and scoring modules. Experimental results and post-evaluation considerations are given in Section 5, and conclusions are drawn in Section 6.

2. System architecture

The system that has been used for this evaluation includes 5 main modules, i.e., Voice Activity Detection (VAD), feature extraction, i-vector extraction and PLDA modeling, score normalization, and score combination and calibration. These modules are described in the next sections.

2.1. Voice Activity Detection

Voice Activity Detection is performed by means of a phonetic decoder. The decoder is a hybrid HMM-ANN model trained to recognize 11 language independent phone classes, including silence. More details are given in [2]. The microphone and interview calls were amplified and sub-sampled. Moreover, before applying the HMM-ANN VAD, a preliminary filtering process was performed for reducing cross-talk effects. On telephone and microphone test segments, overlapping regions with high energy level on both channels were detected as potential cross-talk indicators. The validation of the cross-talk hypothesis was obtained by training and comparing the speaker models associated to those regions in the two channels. In case of cross-talk, the regions corresponding to speech activity on the other channel were removed from the target side. For interview, detection of regions of target speaker activity is based on the comparison of the relative energy of the interviewer and interviewee channels over a sliding window. We retain the frames of the target channel having energy greater than the energy of the interviewer channel in the corresponding windows. A smoothing filter is then applied to avoid chopping the target channel into small segments.

2.2. Feature extraction

Five sets of features have been extracted for training the models used in this evaluation, including two with a relatively "small" dimension and three with a relatively "large" dimension. The use of "small" and "large" versions of the same features was motivated by past experience and development results. All the features, summarized in Table 1, are warped by means of short term Gaussianization [3].

The first set (*MFCC-25*) includes 12 Mel Frequency Cepstral Coefficients plus 13 delta cepstrum. The second set of "small" features (*PLP-26*) includes 13 PLP coefficients (p0-p12) and their first order derivatives. The two sets of "large" features (*MFCC-60, PLP-60*) consist of 60 parameters, 20 coefficients and their first and second order derivatives. The fifth set *MFCC-46* has 46 parameters, 19 MFCC coefficients, 19 first order derivatives and 8 second order derivatives. All sets of features were computed with a frame rate of 100 observation vectors per second.

Table 1. *Extracted features.*

Feature type	Feature number	Features	Δ	ΔΔ
MFCC	**25**	c1-c12	Δc0- Δc12	
PLP	**26**	p0-p12	Δp0- Δp12	
MFCC	**60**	c0-c19	Δc0- Δc19	ΔΔc0- ΔΔc19
PLP	**60**	p0-p19	Δp0- Δp19	ΔΔp0- ΔΔp19
MFCC	**46**	c1-c19	Δc0- Δc18	ΔΔc0- ΔΔc7

For the *MFCC-25* and *MFCC-46* sets of features, the analysis bandwidth is 300-3400 Hz. Feature warping to a Gaussian distribution is performed for each static parameter stream on a 3 sec sliding window excluding silence frames. All the other feature sets are extracted analyzing the full 0-4000 Hz bandwidth and feature warping is performed before the VAD has been applied, thus including silence frames.

2.3. Speaker models

The acoustic speaker models for this evaluation are i-vectors [4] extracted from Gaussian Mixture Models (GMMs). The models, consisting of 2048 diagonal Gaussian mixtures, were trained running 10 iterations of an approximation of the EM algorithm. In this approximation, only the best Gaussian statistics within each frame are updated for the sake of efficiency.

The gender independent *MFCC-25* and the gender dependent *MFCC-46* UBMs were trained using the conversations of the NIST SRE 2006, 2008 and 2010 databases of the speakers in the SRE 2012 training list. The training set includes 737 hours of speech selected from the 21780 conversations of 1095 female speakers and 512 hours from 15726 conversations of 723 male speakers. Since we had available gender dependent UBMs with 1024 Gaussians for *PLP-26, MFCC-60,* and *PLP-60* front ends, which were trained for the NIST 2010 evaluation, the new 2048 Gaussian models have been trained by splitting these models and re-estimating their parameters using the SRE 2012 training data.

Maximum-Likelihood estimation of the sub-space matrix $\mathbf{T}$ was obtained by minor modifications of the Joint Factor Analysis approach [5], using the same dataset exploited for training the UBMs. The dimension of the i-vectors has been set to 600 for the 60-feature models and to 400 for the other models, respectively.

2.4. Speaker classification

Gaussian PLDA systems have been used for recognition, implemented according to the framework illustrated in [6], [7]. In particular, we trained PLDA models with full–rank channel factors, using 200 dimensions for the speaker factors. The i–vectors used for the PLDA models are L_2 normalized.

PLDA training was performed using a balanced set of clean and noisy utterances for every speaker. The Filtering and Noise Adding Tool [8] implementing the “C-message” weighting function for SNR estimation has been used for obtaining noisy replicas of clean utterances. In particular the noisy utterances were obtained artificially adding HVAC or crowd noise according to the NIST specifications. Random Signal-to-Noise Ratios in the range of 4 to 17dB were used. Three versions of each development conversation were obtained, namely one that is clean, one contaminated with HVAC noise, and one contaminated with crowd noise. No appreciable improvement on development tests was obtained when training a different PLDA for clean and for noisy speech.

The training and test datasets for development were selected from the SRE 2012 training data of the target models. Here we eliminated the 10sec and the summed conversation utterances. We ensured that highly correlated segments (e.g. same interview from different microphones) were assigned either to the training or to the test set. The development training set finally included 737 hours of speech selected from the 21780 conversations of 1095 female speakers and 512 hours from 15726 conversations of 723 male speakers.

The PLDA of each sub-system was trained randomly selecting one clean or noisy version of each utterance from the development training set. These PLDAs were used for estimating the fusion and calibration parameters. For scoring the evaluation data, a new set of PLDAs was trained including all the SRE 2012 training data. Post evaluation demonstrated the validity of this approach both in term of accuracy and calibration.

3. Development test sets

The common evaluation conditions in the SRE12 evaluation plan defined five subsets of trials in the core test. These satisfy additional constraints including recordings of interviews and of telephone calls, with and without added noise, plus speech segments intentionally collected in a noisy environment. The latter subset, however, has an average SNR comparable with the one of the clean segments, as also can be appreciated by comparing the results of the clean and noisy telephone calls trials. Moreover, three optional test segment conditions were defined: the extended test set, much larger than the core test set, the “Known” subset, which is the set of segments belonging to trained speakers, and the “Unknown” subset, which includes, as impostor trials, segments that do not belong to any trained speaker model.

The 75% partition of SRE 2012 training data was used for training the PLDA, the remaining 25% for estimating the fusion and calibration weights, and as the development test set. This partition was further extended adding two utterances from 100 male and 100 female speakers from SRE05. The aim was to test the accuracy of the models in a condition similar to the “Unknown” evaluation condition. Moreover, the set was extended adding conversation copies with different SNR and durations. This set was also used for training the condition-dependent calibration parameters. Four classes corresponding to the segment duration ranges 4-12 sec, 12-20 sec, 20-40 sec, and 40-60 sec, respectively, were defined to quantify the variability of accuracy due to the speech amount. Additionally, in order to account for tests with different SNR, we included two versions corrupted with HVAC/crowd noise at 6 and 15dB according to the NIST specifications. Finally, to contrast the effect of having few training segments by estimating a tuned calibration on them, we added a set of target speaker models trained with only one and two segments. These speakers were selected among the target speakers having more than 3 training segments, for a total of 1880 and 1212 additional female and male models, respectively.

For speakers having more than one or two segments, the development test sets included 3015 and 2086 true speaker trials for female and male speakers, respectively. The corresponding impostor trials were 328K and 216K, respectively. The addition of models trained with less than 3 segments, allowed us to enlarge the test sets to 5658/3954 true speakers and 564K/360K impostor trials for female and male speakers respectively.

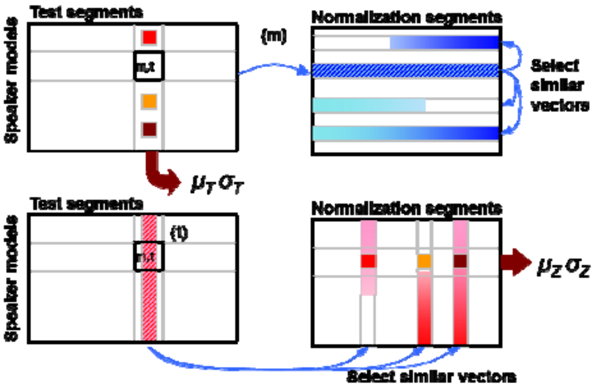


Figure 1: *Simplified Adaptive Symmetrical score Normalization.*

4. Score normalization

The score of each classifier was normalized according to a simplified version of the Adaptive Symmetrical score Normalization (AS-Norm) [9]. The AS-Norm is derived from the AT-Norm [10], but preserves the symmetrical property of the S-Norm [4]. The PLDA score s comparing two i-vectors i_1 and i_2 is normalized according to:

$$\frac{1}{2} \cdot \left[\frac{s - \mu_1(N_2)}{\sigma_1(N_2)} + \frac{s - \mu_2(N_1)}{\sigma_2(N_1)} \right] \qquad (2)$$

where μ_1 and σ_1 are the mean and standard deviation of the scores obtained by matching i-vector i_1 against a normalization subset N_2 depending of i-vector i_2, and the same notation dually applies to the second term in parenthesis. The selection of the normalization subset follows the procedure in [10].

In the simplified AS-Norm implementation, used for SRE12 and shown in Figure 1, the normalization set for the test segments ("T-Norm" component of the "S-Norm"), is the whole set of target models, augmented by the models trained with one and two segments only, for a total of 3035 and 1975 female and male models, respectively. The normalization set for the target voiceprint normalization ("Z-Norm" component of the "S-Norm") is instead obtained selecting a random segment for each target speaker, and adding clean, HVAC and crowd versions to the normalization set. The total number of segments for normalization was 3285 and 2169 for female and male, respectively.

The number of elements for computing the means and standard deviations, selected by the adaptive AS-Norm approach, was set to 100. The combination of the AS-Norm scores of the 5 models described in Section 2.3 is obtained by linear fusion with prior-weighted Logistic Regression objective [11]. Here, we estimated the combination parameters on the development test set (described in Section 3) using the BOSARIS toolkit [12]. Condition dependent calibration parameters are trained on the basis of the segment gender, number of training segments, the test segment duration and SNR, according to the classes given in Table 2, for a total of 80 calibration parameter sets. No distinction has been made between HVAC and crowd noise, but we decided to use 4 classes conditioned to the SNR, for better sampling the accuracy variation as a function of the noise level.

For each test segment, we scored all the target models, regardless the test segment condition (Core, Extended, Known or Unknown) and then we applied the transformation for obtaining the compound LLR, as described in [12]. The transformation was not applied for the 'Unknown' condition.

Table 2. *Calibration classes.*

Category					
Gender	Male	Female			
Training segments N.	<=2	>2			
Test segment duration [sec.]	<12	12-20	20-40	40-60	>60
SNR [dB]	<9	9-14	14-21	>21	

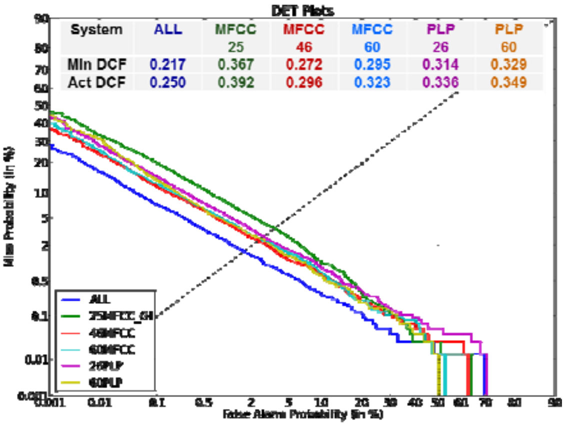
5. Results

The same combination of systems has been used for all the test conditions. Figure 2 summarizes the results of the five subsystems, and their fusion for the *telephone without added noise condition* common conditions (CC2).

Looking at the DET curves, and at the minimum and actual DCFs, it can be seen that when excluding the small Gender Independent *MFCC-25* system, all the others give comparable performance in terms of the DCF defined for the SRE12 evaluation [1]. Significant improvement is obtained by the combination of the subsystems.

Figure 3 illustrates the performance of the fused system for the 5 common conditions defined by NIST, including recordings of interviews and of telephone calls, with and without added noise. The test are performed on 4 test conditions of the official training / test matrix, i.e. Core, Extended, Known and Unknown, using the set of target speaker of the core training condition. Each tuple of bars shows the actual and minimum DCF obtained in the corresponding condition.

On telephone common conditions, we obtained almost perfect calibration when scoring the known subsets. On interviews instead, the best calibration was obtained on extended and unknown test conditions. On average however, the calibration on the core tests seems to be more critical with respect to other test conditions: this is particularly evident in the interview with added noise (common condition 3) where there is a large calibration error on the core test and an almost perfect calibration on the extended condition.



System	ALL	MFCC 25	MFCC 46	MFCC 60	PLP 26	PLP 60
Min DCF	0.217	0.367	0.272	0.295	0.314	0.329
Act DCF	0.250	0.392	0.296	0.323	0.336	0.349

Figure 2: *Telephone without added noise condition CC2.*

Table 3. *Comparison on the extended set tests of a subsystem (MFCC-46) with different normalization and calibration techniques.*

Condition	Tel without noise (CC2)				Tel with noise (CC4)				Tel Noisy(CC5)				Interview without noise (CC1)				Interview with noise (CC3)			
ASNorm	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
Cond-dep Cal	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	No	No
EER	2.42	1.66	2.42	1.84	3.14	4.55	2.87	5.07	2.73	2.06	2.68	2.24	1.97	2.28	2.35	2.51	2.75	3.47	3.06	3.71
Min_DCF100	0.223	0.176	0.218	0.179	0.260	0.281	0.245	0.243	0.261	0.200	0.252	0.201	0,133	0.124	0.141	0.143	0.110	0.106	0.119	0.120
Min_DCF1000	0.359	0.319	0.344	0.302	0.368	0.408	0.366	0.372	0.390	0.343	0.378	0.327	0.201	0.195	0.215	0.219	0.162	0.170	0.176	0.183
Min_DCF12	0.272	0.248	0.281	0.241	0.314	0.344	0.306	0.308	0.326	0.272	0.315	0.264	0.167	0.160	0.178	0.181	0.136	0.138	0.148	0.151
Act_DCF12	0.296	0.328	0.283	0.274	0.354	0.425	0.312	0.337	0.344	0.384	0.323	0.318	0.170	0.167	0.185	0.184	0.137	0.144	0.158	0.159

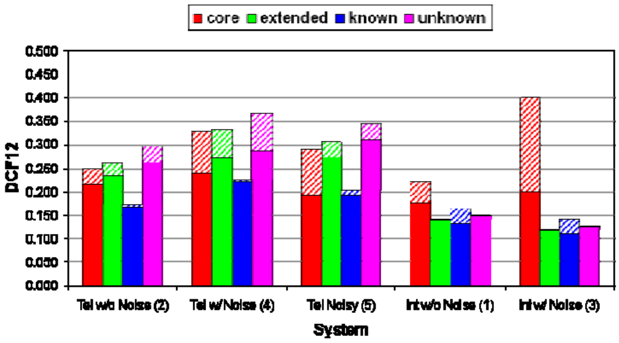


Figure 3: *Min (solid) and Actual (dashed) DCF12*

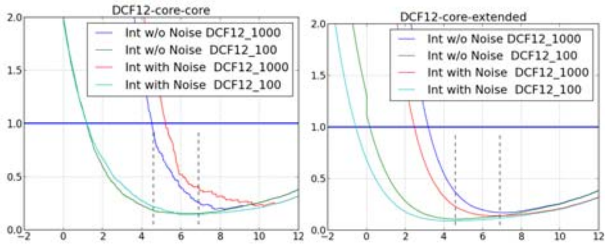


Figure 4: *DCF plot comparison for interview, core-core and core-extended test conditions*

Figure 4 unveils the nature of this phenomenon: the plot on the left refers to the core test condition, whereas the plot on the right refers to the extended test condition. The graphs show the two components of the SRE12 DCF function, with 1/100 and 1/1000 target prior probabilities, for interview without noise (blue and green curves) and interview with noise (cyan and red curves). Comparing the core and extended plots, it can be seen there is a right offset of the DCF curves in the left figure: this means that the calibration data used in development does not perfectly match the actual core test condition. Moreover the blue and the red curves of the left graph are quite stepped, suggesting data sparseness and unreliability.

5.1. Post-evaluation

Table 3 summarizes the results of post-evaluation tests of one of the submitted subsystems (*MFCC-46*), aimed at analyzing the effect of AS-Norm and condition dependent LLR calibration, as used in the evaluation system. In particular the two first lines of the table define the configuration of the experiment, while the remaining lines show the results in terms of Equal Error Rate, Minimum DCF100 and 1000, (with 1/100 and 1/1000 target prior probability, respectively) and Minimum / Actual DCF12 as defined by the SRE12 evaluation plan [1]. The *MFCC-46* subsystem has been selected because it provided the best performance, as a single system, in all conditions. Also it is representative of the fused system behaviors, as can be appreciated comparing the figures in Table 3 with the bars in Figure 3.

When the condition dependent LLR calibration is not used, the calibration accounts for gender dependency only (two classes calibration). The main outcomes can be summarized as follows:

- On the interview conditions (CC 1 and 3), the condition dependent calibration was effective on accuracy. The effect of AS-Norm was relevant on EER, whereas is minimal on DCF 12 (both minimum and actual).
- AS-Norm provides quite good calibration in all conditions (the maximum difference between min and actual DCF12 is ~13% in CC 4), but it was suboptimal on telephone conditions (CC 2, 4 and 5).
- Two classes calibration, with or without AS-Norm, is more effective in actual DCF, than condition dependent calibration, on telephone conditions.
- Best EER on telephone without added noise (CC 2 and 5) is obtained without AS-Norm.

Overall, our SRE12 evaluation does not supply clear evidence of the AS-Norm effectiveness. However, since condition dependent calibration was used for all conditions, AS-Norm was essential on telephone tests for limiting the calibration loss, which otherwise would be in the range 23-40%. As far as condition dependent calibration is concerned, it provided roughly a 10% improvement in interview common conditions but it was not effective on telephone tests.

In conclusion, considering the two extremes, i.e., AS-Norm plus condition dependent calibration with respect to two class calibration alone some advantage has been obtained in the interview conditions, while we lose up to 8% in minimum DCF on telephone conditions

6. Conclusions

We presented the components and analyzed the results of the Nuance–Politecnico di Torino (NPT) speaker recognition system submitted to the NIST SRE12 evaluation campaign. We have shown that the use of AS-Norm plus condition dependent calibration has been successful for the evaluation, although the results do not provide a clear evidence of a wide range effectiveness of the selected score normalization and calibration techniques. In particular, these post-evaluation results are sometimes in contrast with our past experience with AS-Norm, which usually provides appreciable benefits when tuned on data consistent with the target application scenario.

7. References

[1] National Institute of Standards and Technology, “NIST speech group web”, http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf.

[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] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint Factor Analysis versus Eigenchannels in Speaker Recognition”, IEEE Transaction on Audio Speech and Language Processing, vol. 15, no. 4, pp. 1435–1447, 2007.

[6] P. Kenny, “Bayesian speaker verification with heavy–tailed priors”, in Keynote presentation, Odyssey 2010, The Speaker and Language Recognition Workshop, 2010. Available at http://www.crim.ca/perso/patrick.kenny/kenny_Odyssey2010.pdf.

[7] N. Brummer and E. de Villiers, “The speaker partitioning problem”, in ¨Proc. of Odyssey 2010, 2010, pp. 194–201.

[8] H. G. Hirsch, “Filtering and Noise Adding Tool”. Available at http://dnt.kr.hsnr.de/download.

[9] 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.

[10] 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

[11] N. Brummer and J. du Preez, “Application-Independent Evaluation of Speaker Detection” Computer Speech & Language Vol. 20, 2-3, pp. 230-275, 2006.

[12] Available at https://sites.google.com/site/bosaristoolkit/home