TECHNICAL REPORT NO. 1

Technical report for contract between BUT and Phonexia, "výzkum a následný vývoj systému pro rozpoznávání mluvčího"

Pavel Matějka, Oldřich Plchot and Ondřej Novotný

Brno University of Technology, Speech@FIT group

December 15th 2016

1 INTRODUCTION

The goal of the project is to improve robustness and accuracy of the speaker recognition systems. The main problem of the nowadays speaker recognition system is the language and channel variability. It is observed that new language and channel completely unseen during the training of the system produce much worse performance of the system.

National Institute of the Standardization and Normalization (NIST) organized a Speaker recognition evaluation (SRE) 2016 with the similar goals: unknown language, small amount of development data, short files, different channels,…. The results described in this report are on the evaluation data from NIST SRE 2016 evaluation. We performed two main experiments. In the first, we compared different feature extractions and in the second, we analyzed the most important component in the system - adaptive score normalization which gave 30% relative improvement in performance.

2 SPEAKER RECOGNITION SYSTEM

Conventional speaker recognition system is composed of several blocks which are shown in Figure 1. At first, the Feature extraction module which spectral featuress from audio. The second block is iVector extraction [3] which produces low-dimensional vector representation of arbitrary length audio. Such vector is post-processed with length normalization, LDA (dimensionality reduction), and WCCN (within class covariance normalization). Then, two iVectors (enrollment and test) are compared with the Classifier (PLDA – Probabilistic Linear Discriminant Analysis) [4]. Next, calibration/normalization of the scores are applied, this step is crucial because it converts the scores to the meaningful values which are then presented to the user. More description of iVector based system can be found in [1,5].

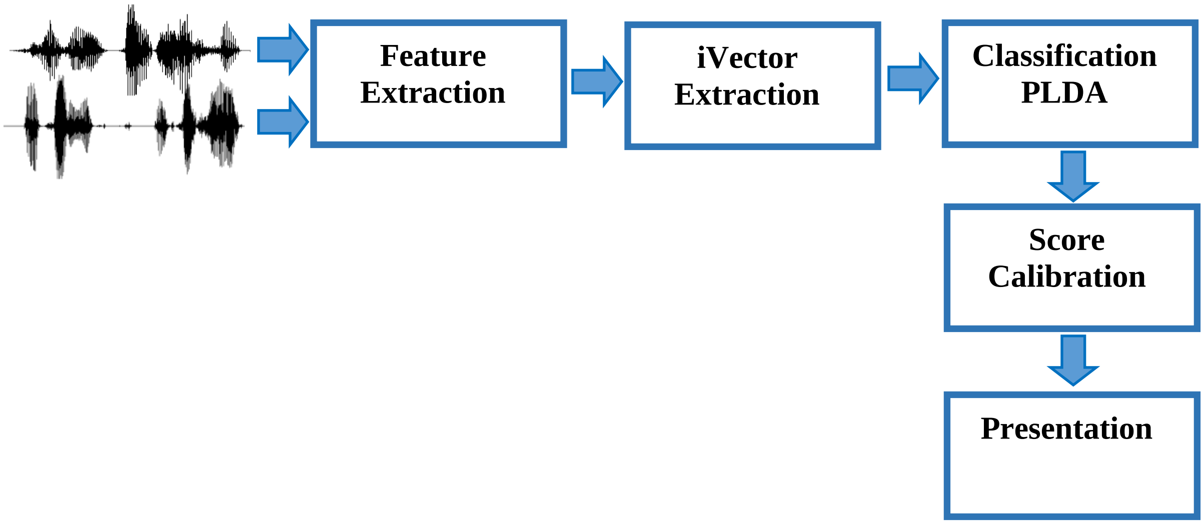

Figure 1: Speaker recognition system

3 SCORE NORMALIZATION AND CALIBRATION

Score normalization is one of the most important part of the system, because it converts the scores to values which are presented to the user. Ideally, we want to have log-likelihood ratio as the output of the system.

If the system is exposed to the data which were not seen during the training, this output might be shifted and scaled and give not valid numbers. The discriminative power of the system might still perfectly work, but the output presented to the user is wrong. The example of such miss-calibrated output is shown in Figure 2. Blue and red lines represents histogram of the non-target and target distributions. If these were well calibrated, their intersection should be above zero.

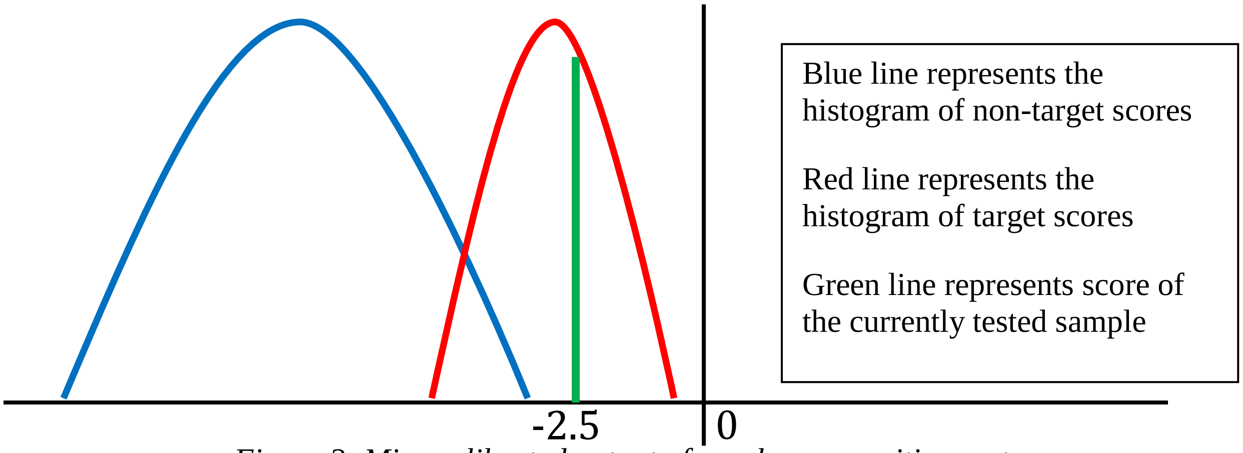

Figure 2: Miss-calibrated output of speaker recognition system

If we do the comparison of enrollment and test samples from the same speaker (target) and get the score with such system (example represented by green line and score value -2.5), we get perfect hit and the discrimination of the system is still perfect, since the green line falls under the red distribution. But the score itself is -2.5 which is below 0, so the decision would be “it is not the same speaker”. The solution for this is to “show” to the system how target and non-target scores look like in this particular domain and do the shift and scale of the score.

This is a task of **score normalization and calibration:** to shift and scale the scores to be well calibrated and represented as log-likelihood ration. Figure 3 shows the same example as the previous one but with well calibrated system.

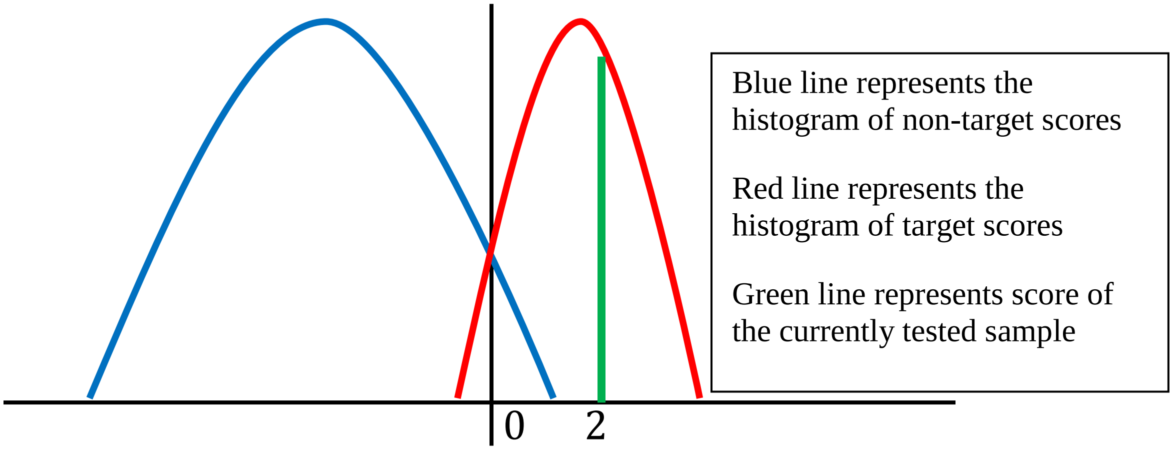

Figure 3: Well calibrated output of speaker recognition system

Score normalization – is non-target based score normalization and the system needs to see lots of non-target scores – hundreds to thousands. It centers the non-target (blue) distribution to have zero mean. The advantage is that for this we do not need to have speaker labels for the target data. But the output is not log-likelihood ration. The calibration still has to be done in the system. But even if we do only the score normalization, we get very good and usable results. The most common score normalization for the iVector based system is so called s-norm (symmetric-norm). In s-norm, we have a normalization cohort (set of utterances ideally from the target channel) and enrollment and test utterance which we want to score. To get the normalized score we:

1. Score test and enrollment against each other and get un-normalized score.

2. Score enrollment sample to all files in normalization cohort and compute mean and variance of all these scores. Divide the score from 1) by this variance and subtract mean.

3. Score test sample to all files in normalization cohort and compute mean and variance of all these scores. Divide the score from 1) by this variance and subtract mean.
4. Final score is computed as an average of output from 2) and 3)

The extension of this method is an adaptive score normalization. The difference is that in 2) and 3), we do not take all values to compute mean and variance but we sort the scores and take X top scoring values for mean and variance computation, where X might be from 100 to 1000 and needs to be experimentally tuned.

Score calibration – learns the mean and shift of the target and non-target distribution. The output is log-likelihood ratio. We need to have database of the speakers from target domain with speaker labels, which is usually hard to get in real scenarios.

4 NIST SRE 2016 DATASET[1]

This evaluation simulates a real word scenario where enrollment segments contain up to 60 seconds of audio and test segments contain between 10-60 seconds of audio. The actual data come from different languages and channels than we have seen in previous NIST evaluation.

Data - The data collected by the LDC as part of the Call My Net Speech Collection to support speaker recognition research were used to compile the SRE16 test, development and training sets. The data are composed of telephone conversations collected outside North America, spoken in Tagalog and Cantonese (referred to as the *major languages*) and Cebuano and Mandarin (referred to as the *minor languages*). The development set described below contains data from the two minor languages, while the test set contains data from the two major languages. NIST released the development set that mirrors the evaluation conditions. The development data are drawn from the minor languages and include: 20 speakers, 10 each from Cebuano and Mandarin, 10 calls per speaker + associated metadata (speaker id, gender, language, and phone number). There is also a set for unsupervised training that contains 2000 calls from major languages and 200 calls from minor languages without any further annotation – we call it *unlabeled data*.

[1] Evaluation plan of the NIST SRE 2016
https://www.nist.gov/sites/default/files/documents/itl/iad/mig/SRE16_Eval_Plan_V1-0.pdf

There are two types of training conditions - **Fixed** (with predefined training data) and **Open** (with any publicly available data from LDC/ELRA) + 11 languages from BABEL program released for this evaluation.

Enrollment condition is defined as the number of speech segments provided to create a target speaker model. However, unlike previous SREs, gender labels are not provided. The duration of speech is around 60 seconds. There are two conditions:

- One segment - the system is given only one segment to build the model of the target speaker
- Three-segment – the system is given three segments to build the model of the target speaker, all from the same phone number

Test condition - Speech durations of the test segments are uniformly sampled ranging from 10 secs to 60 secs. Trials are conducted with test segments from both same and different phone numbers as the enrollment segment(s). There are no cross-sex trials. There are no cross-language trials.

4.1 System description

The topology of the system is conventional iVector based system [1] with 2048 Gaussian Mixture Models, iVector with 600 dimensions followed by LDA dimensionality reduction to 200 components. The classifier is PLDA. The training data comes from previous NIST SRE recognition evaluation. Particularly, we used a division which we proposed for modeling also multilingual dependences, noise and reverberation issues. The set is PRISM evaluation set described and downloadable from [2]. The initial results on PRISM set indicating that language variability is a problem are described for several systems in [1].

4.2 Comparison of conventional acoustic features and bottleneck features

We performed a comparison of 5 systems with different feature extraction. Three of them are purely acoustic features (MFCC, PLP and Perseus) and 2 of them are the best performing system architecture developed prior to NIST SRE 2016. This systems are a feature level fusion of MFCC with Bottleneck features and give about 50% relative improvement in accuracy on English data (NIST SRE 2010). The conclusion is that all systems perform about the same. There is no dramatic improvement on this data when the Bottleneck features are present in the system.

Features	feadim	EER[%]	minCprim
MFCC	60	13.13	0.7535
PLP	60	13.08	0.7578
MFCC+SBN80-BABEL (open cond)	140	13.12	0.7727
Perseus	60	13.68	0.8125
MFCC+SBN80-Fisher	140	14.98	0.8174

4.3 Adaptive score normalization

Adaptive score normalization was the essential step in the successful system for this evaluation across all sites who submitted the system. Figure 4 shows the improvement from the normalization and adaptation techniques we tried. The technique which gives the most of the improvement from the baseline is the *adaptive score normalization* (all other techniques give smaller improvement). Further to the left of the figure are other techniques which did not add any improvement on the evaluation data.

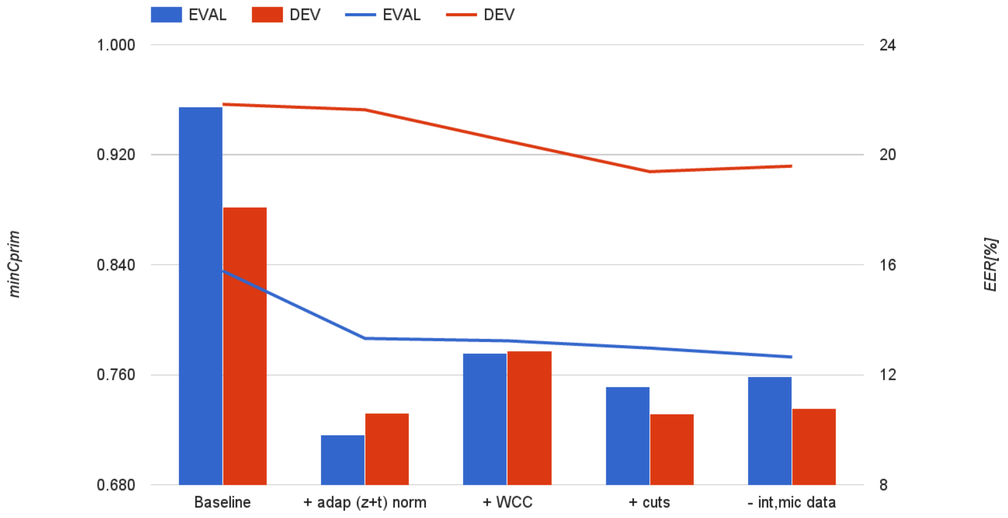

Figure 4: Effects of adaptation

An analysis of the score normalization is shown in Figure 5. There is a comparison of the several normalization techniques: z-norm, t-norm and s-norm. For our normalization cohort, we used all our training data (~70k files) together with DEV16 unlabeled data (~2k files, development data released by NIST for this evaluation, this data contain similar recordings to the evaluation data, target channel data). The first part of the graph uses all data from the normalization cohort and shows that s-norm is slightly better than other normalization techniques. If we use all data only from DEV16 unlabeled, we observe dramatic improvement.

The second half of the graph shows the same normalization techniques but using adaptive selection of the normalization cohort. This mean that for each enroll and test utterance, we select 1000 best scoring utterances from the normalization cohort to compute the normalization parameters (mean and variance). There is again a big improvement in all normalization techniques; S-norm produces the best results. When doing adaptive s-norm, there is almost no difference in the results if we use only target channel data in normalization cohort. Deeper analysis shows that when using all normalization data in the cohort, then in average there is about 50% of the data from DEV16 unlabeled data. We also run the experiments where we varied the number of selected top scores from the cohort and with top-100, we reached the same results as with 1000.

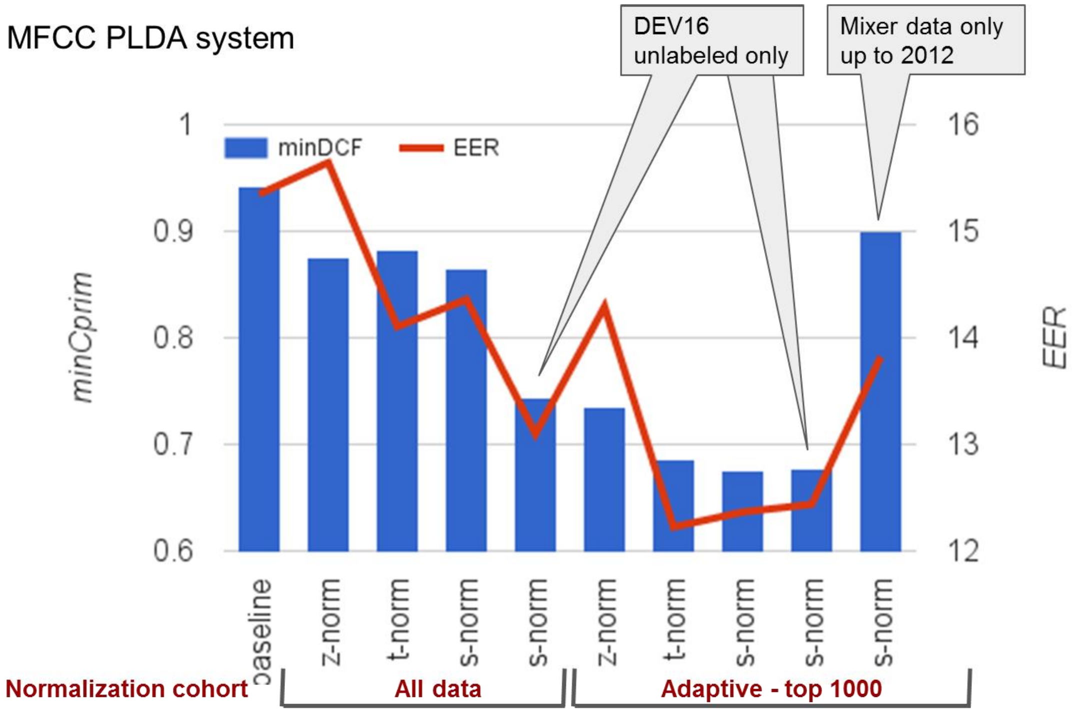

Figure 5: Effects of score normalization

The Last column in the graph shows also the adaptive normalization technique but without DEV16 unlabeled data. The result is better than baseline but much worse than if we use the DEV16 unlabeled data.

The final message is that we need to use adaptive score normalization and there has to be data from the target channel in the normalization cohort.

5 CONCLUSION

We presented the analysis of the system behavior for unseen language and channel. The results are presented on the NIST SRE 2016 data where the system was trained mostly on English data and the test data are from Tagalog and Cantonese. We can observe big performance degradation. Unfortunately it is not possible to have directly comparable numbers. But on NIST SRE 2010 English only scenario with 2 minutes of speech per file the system performance (EER) is between 1-3%. Our development set for NIST SRE 2016 consist of Mandarin data (recorded in USA) with 10-60 seconds of speech has EER in the range of 7-10%. Initial single best results on the NIST SRE 2016 data (Tagalog and Cantonese) recorded in Asia with 10-60 seconds of speech per file is EER=20%.

We have experimented with different adaptation and normalization techniques. The best performing was the adaptive score normalization with Tagalog and Cantonese development data provided by NIST in the normalization cohort. This decreased the EER to 16%. Further fusion of 4 systems decreased the EER to 13%.

Further analysis is needed to investigate where the errors come from, if this is from duration, language, channel, etc. We will use these findings to improve our adaptation strategies.

6 REFERENCES

[1] NOVOTNÝ Ondřej, MATĚJKA Pavel, GLEMBEK Ondřej, PLCHOT Oldřich, GRÉZL František, BURGET Lukáš and ČERNOCKÝ Jan. Analysis of the DNN-Based SRE Systems in Multi-language Conditions. In: *Proceedings of SLT 2016*. San Diego: IEEE Signal Processing Society, 2016, pp. 199-204. ISBN 978-1-5090-4903-5. http://www.fit.vutbr.cz/research/groups/speech/publi/2016/novotny_slt2016_0000199.pdf

[2] 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

[3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. PP, no. 99, 2010

[4] 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.

[5] MATĚJKA Pavel, GLEMBEK Ondřej, CASTALDO Fabio, ALAM Jahangir, PLCHOT Oldřich, KENNY Patrick, BURGET Lukáš and ČERNOCKÝ Jan. Full-covariance UBM and Heavy-tailed PLDA in I-Vector Speaker Verification. In: *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011*. Praha: IEEE Signal Processing Society, 2011, pp. 4828-4831. ISBN 978-1-4577-0537-3.