CALIBRATION BASED ON DURATION QUALITY MEASURES FUNCTION IN NOISE ROBUST SPEAKER RECOGNITION FOR NIST SRE'12

Miranti Indar Mandasari, Rahim Saeidi, David A. van Leeuwen

Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands

ABSTRACT

This paper presents the performance evaluation of i-vector and PLDA based speaker recognition system which incorporate *quality measures function* (QMF) in linear calibration. Evaluated on the recent NIST SRE'12 corpus, the linear calibration with QMF as the additional term shows a positive gain in the system performance compared to the conventional linear calibration with only two terms. Based on the equal error rate values measured from our I4U evaluation trial set, the QMF calibration approach shows 10 – 37 % relative improvement compared to the conventional linear calibration. It is shown that by adding 1 – 2 extra parameters in the linear calibration through QMF approach, there is a potential to improve the calibration and discrimination performances of a speaker recognition system.

Index Terms— Calibration, duration, quality measures, QMF, speaker recognition.

1. INTRODUCTION

In a specific field such as forensic speaker identification [1], calibration is very important in order to make the scores produced by an automatic speaker recognition system more reliable. By presenting calibrated scores in the likelihood ratio form, results can be used as legal evidence in court [2]. In the more general field of speaker recognition, the significance of calibration is becoming more recognized by the community. Especially in the 2012 edition of the speaker recognition evaluation (SRE) from the National Institute of Standard and Technology (NIST), calibration is an interesting topic to be discussed amongst the researchers in the field [3]. This is because of the requirement by NIST to participants to express their system output in the log-likelihood-ratio form.

The NIST SRE'12 provides plenty of challenges to its participants. In terms of quality measures of speech samples, there are two problems addressed in this year evaluation, duration variation (20–160 s) and noisy speech conditions. This year's evaluation also includes new performance measure called *primary cost* that is defined as the average of Bayes error rates from two detection cost functions. To address some of the interesting problems offered by the SRE'12 evaluation, this paper presents the performance analysis from calibration that is based on the duration quality measures function (QMF) approach.

The duration QMF for calibration in speaker recognition system is proposed in [4]. In this approach, we add the QMF as an extra term in the linear calibration. Evaluated on the NIST SRE'08 and SRE'10 corpora with truncations to shorter duration, it has been observed that QMF calibration is robust in the conditions where speech duration varied. In this paper, motivated by the challenging NIST SRE'12 protocol, we use the QMF approach as a simple yet robust technique to deal with the duration variability effect on the discrimination and calibration performance of speaker recognition system.

This paper presents the speaker recognition system and corpora descriptions in Section 2 and Section 3, respectively. Calibration methods are detailed in Section 4 and performance measures are listed in Section 5. Experiment results are discussed in Section 6. This paper is concluded in Section 7.

2. SPEAKER RECOGNITION SYSTEM

System configuration of speaker recognition in this paper is fairly similar with the configuration in the latest papers of the authors [4,5]. The system is based on i-vector [6] framework and probabilistic linear discriminant analysis (PLDA) modeling [7, 8]. The main difference is that there is an inclusion of speech enhancement algorithm with a *dynamic noise suppression* rule [9] in the system used for this paper. For this noise suppression purpose, we did noise estimation through *improved minima controlled recursive averaging* (IMCRA) [10], and Wiener filter is applied on the amplitude spectrum as a soft mask.

Spectral features used in this system is 60 dimensional MFCCs which consist of 19 base MFCCs and log energy, augmented with deltas and double deltas. The features are extracted every 10 ms using 30 ms window. To enhance the features quality, a feature-warping was applied [11]. A speech activity detection (SAD) is implemented to extract the active speech frames from the features [12]. Gender-dependent with 2048 components UBM (universal background model) was trained using NIST SRE 2004–2006, Switchboard Cellular phase 1 and 2, and Fisher English corpora.

The i-vectors were trained using a low dimensional (400 dimensions) matrix that defines both the speaker and channel subspaces. Linear discriminant analysis (LDA) projection was applied in order to reduce the i-vectors dimension to 200. Prior to PLDA modeling, the i-vectors were processed by i-vector centering, within-class covariance normalization (WCCN) and length-normalization.

3. CORPORA

The speaker recognition system calibration performance is evaluated using NIST SRE'12 corpus. There are three different datasets used in the experiments. The first two is the *Dev-I4U* (development) and *Eval-I4U* (evaluation) sets from I4U[1] trials list [13]. The third dataset used in the experiments is the evaluation set from NIST SRE'12 trials list which refers to *Eval-SRE'12* set in this paper. For this NIST SRE 2012 evaluation, we made three sub-selections of the core-test core-training condition for different noise levels, based on the 5 common conditions (cc's) defined in the evaluation plan, using

The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement number 238803.

[1] *I4U* is a joint effort from 9 research *I*nstitutes and *U*niversities across *4* continents in joining the NIST SRE'12 evaluation. The lists are available via `http://lands.let.ru.nl/~saeidi/I4U.tgz`

Table 1. Number of trials in the Dev-I4U, Eval-I4U and Eval-SRE'12 sets for female gender and "unknown" non-target trials.

Set	Noise	Number of Trials	
	condition	Target	Non-target
Dev-I4U	Clean	6621	2118521
	15 dB noisy	6621	2118521
	6 dB noisy	6621	2118521
Eval-I4U	Clean	6921	2997225
	15 dB noisy	6921	2997225
	6 dB noisy	6921	2997225
Eval-SRE'12	Clean	4353	120223
	15 dB noisy	2913	7908
	6 dB noisy	2912	7908

version 1 of the trial key. The first level, "noise", consisted of all telephone and microphone speech, without noise addition, that were not recorded in a noisy environment (the intersection of cc1 and cc2 with trials from cc5 removed). For the two noisy conditions, "15 dB" and '6 dB," we selected trials from cc3 and cc4 with added noise of types "babble" and "HVAC" at 15 dB and 6 dB, respectively.

We divided each dataset into 3 different subsets based on the noise conditions in the *test segments* of the trials listed as *clean* (no-alteration), *15 dB* and *6 dB* noise-levels subsets. The number of trials for target and non-target scores are presented in Table 1. Partitioning the results with respect to SNR is intended for analysis of the calibration sensitivity with duration function to SNR-levels in test segments. In the training of calibration parameters, the scores were pooled without noting the SNR-levels. Only the *unknown* non-target[2] trials are included in the experiments, and we focus our experiments on female speakers. By looking at the durations of utterances in the NIST SRE'12 database (a histogram is provided in Figure 2), we see there is a high variability in duration, therefore performing consistent calibration is a very challenging task.

4. CALIBRATION

All calibrations performed for the experiments in this paper are based on the linear transformation of scores into calibrated log likelihood ratio scores. There are two calibration approaches used. The first approach is conventional linear calibration with two parameters, and the second approach is linear calibration that applies *quality measures function* (QMF) as an extra linear term in calibration.

4.1. Linear calibration

In the linear calibration, we transform a set of *raw scores s* which produced from the speaker recognition system to a set of *calibrated scores ℓ* using a linear transformation

$$\ell = w_0 + w_1 s, \tag{1}$$

where w_0 is the offset/gain parameter and w_1 is the scaling parameter of calibration. In this paper, this two parameterized linear calibration is referred to as *conventional* calibration approach.

In the experiment, calibration parameters are trained in a set of scores, and then applied to another set of scores to be evaluated. In this paper, the calibration parameters are trained in the Dev-I4U set (including all noise condition subsets) and applied to all sets which are Dev-I4U, Eval-I4u and Eval-SRE'12. The parameters for both conventional and QMF calibration approaches were trained via *logistic regression* [14] using FoCal toolkit[3].

Table 2. Duration quality measure functions (QMFs) proposed for calibration on various duration conditions.

n	QMF: $Q_n(d_m, d_t, \ldots)$	Additional parameters
1	$Q_1 = w_2 \left\lvert \log \dfrac{d_m}{d_t} \right\rvert$	w_2
2	$Q_2 = w_2 \log^2 \dfrac{d_m}{d_t}$	w_2
3	$Q_3 = w_2 \log \dfrac{d_m}{d_c} \log \dfrac{d_t}{d_c}$	w_2, d_c
4	$Q_4 = w_2 \log \dfrac{d_m}{d_c} \log \dfrac{d_t}{d_c} +$ $w_3 \left(\log^2 \dfrac{d_m}{d_c} + \log^2 \dfrac{d_t}{d_c} \right)$	w_2, w_3, d_c

4.2. Quality Measure Function

Quality measures function or QMF calibration is proposed by the authors in [4] and it was analyzed for the SRE'08 and SRE'10 corpora. This calibration approach basically is a linear calibration technique with several extra parameters in the linear transformation. It also includes the quality measures of speech utterance in the calibration, in this case, the duration of active speech. The QMF calibration approach is applied via linear scores transformation that can be formulated as:

$$\ell = w_0 + w_1 s + Q(d_m, d_t, w_2, \ldots) \tag{2}$$

with $Q(d_m, d_t, w_2, \ldots)$ as the function that defines quality measures we use for calibration, and d_m and d_t as the duration of active speech (after SAD) in the model and test segments, respectively. There were multi-sessions enrollment in the NIST SRE'12, thus we use sum of the duration of utterances in model segments as d_m .

There are four QMFs proposed in [4] and all of this QMFs are analyzed in this paper as well. The four QMFs are presented in Table 2. All QMFs are modeled from the behavior of the calibration parameters of linear score transformation (scaling parameters) in various duration conditions of the model and test segments. Figure 1 depicts the behavior of the scaling parameters across duration of model and test segments. The first two QMFs, Q_1 and Q_2 are formed in order to model the large deviation of the scaling parameters when there is a large difference (mismatched) between the model and test segments. The last two QMFs, Q_3 and Q_4 are modeled from the saddle-plane like of the scaling calibration parameters in calibration which is presented in Figure 1 with $d_c = 20$ s.

5. PERFORMANCE MEASURES

There are five performance measures used to characterized the speaker recognition system performance of discrimination and classification, namely equal error rate ($E_=$), primary cost from NIST SRE'12 (C_{primary}), cost of log likelihood ratio calibration (C_{llr}), minimum C_{llr} ($C_{\text{llr}}^{\text{min}}$), and the miscalibration cost (C_{mc}).

5.1. Equal Error Rate

Equal error rate or $E_=$ is the error rate of a binary-classifier when the probability of the false-acceptance rate and false-rejection rate is equal at a certain point in the detection error trade-off (DET) curve. The $E_=$ was computed using `sretools` analysis package[4] in R using relative operating point convex hull (ROC-CH) approach.

[2] This is done for compatibility results with earlier SRE protocols. [3]

[3] Software is available at `https://sites.google.com/site/nikobrummer/focal`

[4] Software is available at `https://sites.google.com/site/sretools/`

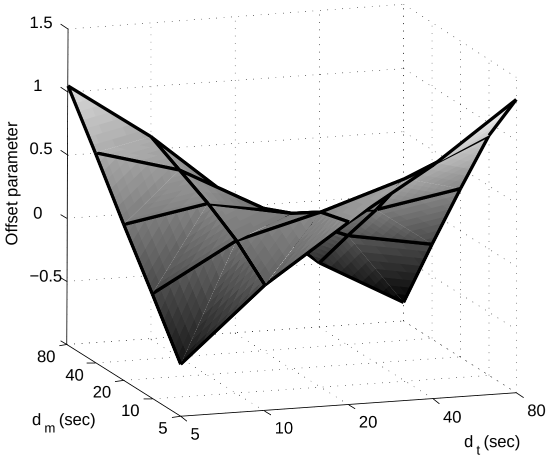

Fig. 1. The saddle-plane shape of calibration (*offset*) parameters in various duration conditions on the model and test segments [4].

5.2. Primary cost for the NIST SRE'12

For this year's speaker recognition evaluation, NIST announced new detection cost function or C_{det} that is referred to C_{primary}. Unlike the previous evaluations, the SRE'12 cost function is a combination of two costs, the cost of NIST SRE'10 ($P_{\text{tar}} = 1/1000$) and another cost with greater prior than SRE'10 ($P_{\text{tar}} = 1/100$). Each of these C_{det} is computed using

$$\begin{aligned} C_{\text{det}} =& C_{\text{miss}} \times P_{\text{tar}} \times P_{\text{miss|tar}} \\ &+ C_{\text{FA}} \times (1 - P_{\text{tar}}) \times (P_{\text{FA|non,known}} \cdot P_{\text{known}} \\ &+ P_{\text{FA|non,unknown}} \cdot (1 - P_{\text{known}})) \end{aligned} \tag{3}$$

with $C_{\text{miss}} = C_{\text{FA}} = 1$. Because in the experiments, we used only unknown non-target trials[5], Equation (3) becomes

$$C_{\text{det}} = C_{\text{miss}} \times P_{\text{tar}} \times P_{\text{miss|tar}} + C_{\text{FA}} \times (1 - P_{\text{tar}}) \times P_{\text{FA|non}}. \tag{4}$$

The C_{det} values are computed using BOSARIS[6] toolkit via Bayes error rate computation.

5.3. Cost of Log Likelihood Ratio Calibration (C_{llr})

As the calibration performance measures, we use cost of likelihood ratio calibration or C_{llr} [15]. The metric C_{llr} can be empirically computed by

$$C_{\text{llr}} = \frac{1}{N_{\text{tar}}} \sum_{i \in \text{tar}} \log_2(1 + e^{-\ell_i}) + \frac{1}{N_{\text{non}}} \sum_{j \in \text{non}} \log_2(1 + e^{\ell_j}) \tag{5}$$

with ℓ_i and ℓ_j as the calibrated log likelihood ratio scores for the target and non-target trials, respectively. Besides C_{llr} , we also use two other measures for calibration namely $C_{\text{llr}}^{\text{min}}$ or the minimum value of C_{llr} and C_{mc} or mis-calibration cost which is defined as the difference between C_{llr} and $C_{\text{llr}}^{\text{min}}$. The metric $C_{\text{llr}}^{\text{min}}$ was computed using *isotonic regression* through pool adjacent violators (PAV) algorithm [16].

6. RESULTS

Results of the calibration experiments conducted in this paper are presented in Table 3. Generally in all evaluated datasets, the system tends to perform slightly better in 15 dB noise condition or 6 dB noise condition (in C_{mc} measure). This is because each of the original NIST segments which are included in the I4U lists has two noise variants included in the training of PLDA and enrollment data. The system is biased to perform better in the slightly noisy conditions compared to the unaltered (clean) condition.

The details of results analysis in Table 3 are divided into two parts: the Dev- and Eval-I4U sets, and Eval-SRE'12 set. Those analysis are discussed in the following.

6.1. Results on I4U Trials List

In this subsection, we present the analysis of calibration experiment results on the Dev-I4U and Eval-I4U sets. In the Dev-I4U set results, one can observe from Table 3 that all QMF calibrations give lower values across all performance measures than the linear calibration on all noise subsets. This is expected because when we applied the calibration parameters trained on the Dev-I4U set to the Dev-I4U set itself (*self-calibration*).

In the Eval-I4U set results, the QMF calibrations perform well across all performance measures when we compared it to the linear calibration results. Table 3 shows that all the QMF calibrations outperform the linear calibration in terms of $C_{\text{llr}}^{\text{min}}$ and $E_{=}$ performance measures. Based on these two performance metrics, Q_1 appears to be the best QMF that provides the best discrimination performance compared to the linear calibration and all other QMF calibrations. The Q_1 calibration results in *absolute* reduction of 0.28 %, 0.47 %, and 0.82 % in the $E_{=}$ compared to the conventional linear calibration on the clean, 15 dB and 6 dB conditions, respectively. This equals to 10 – 37 % *relative* improvement in performance.

In the mis-calibration cost or C_{mc} metric, the QMFs calibration only perform better than the linear calibration in the clean condition. Even though the C_{mc} values for the 15 dB and 6 dB noise conditions for the QMF calibrations are not lower than the linear calibration, still the C_{llr} and $C_{\text{llr}}^{\text{min}}$ values of QMF calibrations are already better than the conventional linear calibration. Using the C_{llr} and C_{primary} measures, the QMF calibrations perform better than the linear calibration in general, with the Q_1 and Q_3 performances slightly surpass the Q_2 and Q_4 calibrations. Evaluated on the Eval-I4U set, the QMF calibrations offer better performance than the conventional linear calibration based on the observations from all five performance measures.

Comparing all four QMFs for calibration, Table 3 shows that the Q_4 performs the best when calibrations applied in the Dev-I4U set while Q_1 performs the best in the Eval-I4U set. This results indicate that the more complex Q_4 function that model the saddle-plane of calibration parameters distribution does not necessary generalize better than the more simple function such as Q_1. The Q_4 training has clearly over-fitted to the calibration development set (Dev-I4U). On the other hand, the simple Q_1 function can be easily and effectively implemented in the *cross-calibration* [7]. Regardless of which QMF is the best for calibration, all QMF calibrations indicate better performances in terms of discrimination and calibration when it is compared to the case where duration information are dismissed.

6.2. NIST SRE'12 Evaluation Results

The experimental results on the evaluation set from NIST SRE'12 (Eval-SRE'12) are slightly different from results on the Dev-I4U and Eval-I4U sets. As presented in Table 3, the QMF calibrations only surpass the linear calibration performance in the 6 dB noise condition subset. In other noise subsets, applying QMF calibrations does

[5] This corresponds to $P_{\text{known}} = 0$ for our experiments.

[6] Software is available at `https://sites.google.com/site/bosaristoolkit/`

[7] Applying the calibration parameters which were trained on one set to another set, in this case, from the Dev-I4U set do the Eval-I4U set.

Table 3. System performance in terms of C_{llr} , $C_{\mathrm{llr}}^{\mathrm{min}}$, C_{mc} , $E_{=}$ and C_{primary} on the Dev-I4U, Eval-I4U and Eval-SRE'12 sets.

Set	Noise cond.	Calibration Method N.A.*	 O**	 Q1	 Q2	 Q3	 Q4
Cost of log-likelihood ratio calibration (C_{llr})							
Dev I4U	Clean	4.373	0.195	0.183	0.183	0.192	**0.178**
	15 dB	2.918	0.078	0.070	0.070	0.071	**0.069**
	6 dB	6.100	0.115	0.100	0.099	0.103	**0.098**
Eval I4U	Clean	3.045	0.170	**0.148**	0.157	0.161	0.172
	15 dB	1.713	0.082	0.078	0.087	**0.072**	0.110
	6 dB	4.338	0.104	0.089	0.098	**0.088**	0.117
Eval SRE'12	Clean	11.099	**0.194**	0.300	0.601	0.306	0.505
	15 dB	5.199	**0.133**	0.183	0.199	0.145	0.256
	6 dB	8.310	**0.179**	0.212	0.232	0.180	0.279
Minimum value of C_{llr} ($C_{\mathrm{llr}}^{\mathrm{min}}$)							
Dev I4U	Clean	0.134	0.134	0.130	**0.129**	0.130	**0.129**
	15 dB	0.066	0.066	**0.057**	**0.057**	0.058	0.058
	6 dB	0.102	0.102	0.089	0.088	0.092	**0.087**
Eval I4U	Clean	0.113	0.113	**0.102**	0.105	0.106	0.104
	15 dB	0.052	0.052	**0.034**	0.036	0.037	0.039
	6 dB	0.086	0.086	**0.057**	0.061	0.065	0.064
Eval SRE'12	Clean	**0.163**	**0.163**	**0.163**	0.266	0.188	0.244
	15 dB	**0.119**	**0.119**	0.129	0.135	0.122	0.145
	6 dB	**0.163**	**0.163**	0.173	0.180	0.165	0.194
Mis-calibration cost (C_{mc})							
Dev I4U	Clean	4.239	0.061	0.054	0.054	0.062	**0.050**
	15 dB	2.852	0.013	0.013	0.013	0.013	**0.011**
	6 dB	5.998	0.014	**0.011**	**0.011**	**0.011**	**0.011**
Eval I4U	Clean	2.932	0.057	**0.046**	0.051	0.055	0.068
	15 dB	1.661	**0.029**	0.044	0.051	0.036	0.071
	6 dB	4.251	**0.018**	0.032	0.037	0.023	0.052
Eval SRE'12	Clean	10.936	**0.031**	0.137	0.335	0.118	0.262
	15 dB	5.080	**0.014**	0.054	0.064	0.023	0.111
	6 dB	8.147	**0.016**	0.039	0.052	**0.016**	0.085
Equal error rate ($E_{=}$) in %							
Dev I4U	Clean	3.43	3.43	3.34	3.33	3.33	**3.32**
	15 dB	1.65	1.65	**1.35**	**1.35**	1.40	1.36
	6 dB	2.53	2.53	2.25	2.24	2.33	**2.17**
Eval I4U	Clean	2.78	2.78	**2.50**	2.53	2.55	2.56
	15 dB	1.27	1.27	**0.80**	0.81	0.86	0.93
	6 dB	2.21	2.21	**1.39**	1.48	1.57	1.64
Eval SRE'12	Clean	**4.22**	**4.22**	4.36	7.19	5.01	6.43
	15 dB	**2.85**	**2.85**	3.15	3.31	2.88	3.67
	6 dB	4.14	4.14	4.40	4.47	**4.12**	5.11
Primary cost for NIST SRE'12 (C_{primary})							
Dev I4U	Clean	0.219	0.219	0.173	0.177	0.205	**0.171**
	15 dB	0.155	0.155	0.153	0.154	**0.152**	0.163
	6 dB	0.249	0.249	**0.235**	0.240	0.236	0.252
Eval I4U	Clean	**0.174**	**0.174**	0.204	0.381	0.236	0.382
	15 dB	0.148	0.148	0.135	0.137	**0.133**	0.160
	6 dB	0.254	0.254	**0.205**	0.215	**0.205**	0.258
Eval SRE'12	Clean	**0.393**	**0.393**	0.485	1.000	0.693	1.000
	15 dB	**0.340**	**0.340**	0.371	0.377	0.343	0.411
	6 dB	0.456	0.456	0.484	0.494	**0.451**	0.533

* N.A. : *Not applicable or no-calibration performed*
** O : *Conventional linear calibration using w_0 and w_1*

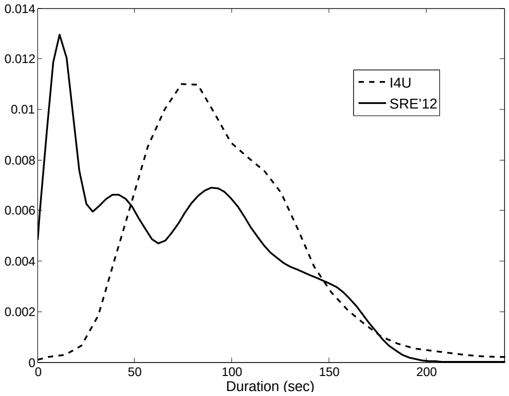

Fig. 2. Distributions of active speech duration from utterances in the I4U and NIST SRE'12 trials.

not seem to give better performance than the linear calibration. To better understand why this is happening, we had a look into the duration distributions of I4U and NIST SRE'12 segments in more details.

In Figure 2, the duration distributions of utterances included in the I4U and NIST SRE'12 lists are depicted. The durations in the plot is the durations of active speech samples for each utterances after the SAD applied. As can be seen from Figure 2, there is quite a difference between the duration distribution of utterance in the I4U and NIST SRE'12 trials lists. The duration distribution of I4U trials list is more concentrated with mean around 90 s of active speech, while the distribution of NIST SRE'12 trials list is more distributed across all duration with a lot of weight in the short duration region.

The difference in range and distribution of duration between the development and Eval-SRE'12 set may be a cause for the QMFs not working very well, but more likely, the 'data set shift' that occurs with every NIST evaluation may be the most important reason. Indeed, the absolute error rates have gone up strongly from Dev-I4U and Eval-I4U to Eval-SRE'12. The subtle changes to the calibration that the QMFs try to make may be lost in the dramatic changes that take place when the data set changes as drastically as it did in going from SRE'10 to SRE'12—despite the fact that all speakers were known in advance. We have some hope, however, that we will be able to get calibration more in line with the SRE'12 material by looking at other quality factors as well.

7. CONCLUSION

Using our development set Dev-I4U to calibrate both the development (self-calibration) and our evaluation set Eval-I4U, the QMF calibration approaches provide a significant performance improvement in both discrimination and calibration. This is observed in all performance metrics used to measures the performance. By adding one or two extra parameters in calibration via the QMF approaches, the system performance based on $E_{=}$ improves by 37 % relative to linear calibration without QMF. However, from the calibration results on the Eval-SRE'12 set using $P_{\mathrm{known}} = 0$, this does not hold. We surmised that in applying a QMF, it is important that the development set matches the evaluation set in terms of duration range and distribution, so that it can give a positive improvement in the system performance. From the problem revealed by the different duration conditions in the Eval-SRE'12 set from the Dev-I4U set for training calibration, our future works include the *truncation* utterances from the development set or simulate the duration effect such that it can be used to better model the duration distribution in the SRE'12 evaluation set.

8. REFERENCES

[1] A. Neustein and H. A. Patil, *Forensic speaker recognition*, Springer, 2011.

[2] J. Gonzalez-Rodriguez and D. Ramos, “Forensic automatic speaker classification in the coming paradigm shift,” *Speaker Classification I*, pp. 205–217, 2007.

[3] National Institute of Standards and Technology, *The NIST Year 2012 Speaker Recognition Evaluation Plan*, Available at http://www.nist.gov/itl/iad/mig/sre12.cfm.

[4] M. I. Mandasari, R. Saeidi, M. McLaren, and D. A. van Leeuwen, “Quality measure functions for calibration of speaker recognition system in various duration conditions,” *Accepted to IEEE Trans. on Audio Speech and Language Procesing*, August 2013.

[5] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, “The effect of noise on modern automatic speaker recognition systems,” in *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4249–4252.

[6] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, “Cosine similarity scoring without score normalization techniques,” in *Proc. of Odyssey, Speaker and Language Recogntion Workshop*. IEEE, 2010, pp. 71–75.

[7] S. J. D. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *Proc. of Int. Conf. on Computer Vision (ICCV)*. IEEE, 2007, pp. 1–8.

[8] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Brümmer, “Discriminatively trained probabilistic linear discriminant analysis for speaker verification,” in *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4832–4835.

[9] R. Saeidi and D. A. van Leeuwen, “The Radboud University Nijmegen submission to NIST SRE-2012,” in *Proc. of the NIST Speaker Recognition Evaluation Workshop*, 2012.

[10] I. Cohen, “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging,” *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[11] J. Pelecanos and S. Sridharan, “Feature warping for robust speaker verification,” in *Proc. of Odyssey, Speaker and Language Recogntion Workshop*. IEEE, 2001, pp. 213–218.

[12] M. McLaren and D. A. van Leeuwen, “A simple and effective speech activity detection algorithm for telephone and microphone speech,” in *Proc. of the NIST Speaker Recognition Evaluation Workshop*, 2011.

[13] R. Saeidi and et. al., “I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification,” in *Proc. of Interspeech*. IEEE, 2013.

[14] S. Pigeon, P. Druyts, and P. Verlinde, “Applying logistic regression to the fusion of the NIST’99 1-speaker submissions,” *Digital Signal Processing*, vol. 10, no. 1-3, pp. 237–248, 2000.

[15] D. A. van Leeuwen and N. Brümmer, “An introduction to application-independent evaluation of speaker recognition systems,” *Speaker Classification I*, pp. 330–353, 2007.

[16] N. Brümmer and J. du Preez, “Application-independent evaluation of speaker detection,” *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.