ANALYSIS OF THE DNN-BASED SRE SYSTEMS IN MULTI-LANGUAGE CONDITIONS

Ondřej Novotný, Pavel Matějka, Ondřej Glembek, Oldřich Plchot,
František Grézl, Lukáš Burget, and Jan "Honza" Černocký

Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Brno, Czech Republic
{inovoton,matejkap,glembek,iplchot,grezl,burget,cernocky}@fit.vutbr.cz

ABSTRACT

This paper analyzes the behavior of our state-of-the-art Deep Neural Network/i-vector/PLDA-based speaker recognition systems in multi-language conditions. On the "Language Pack" of the PRISM set, we evaluate the systems' performance using the NIST's standard metrics. We show that not only the gain from using DNNs vanishes, nor using dedicated DNNs for target conditions helps, but also the DNN-based systems tend to produce de-calibrated scores under the studied conditions. This work gives suggestions for directions of future research rather than any particular solutions to these issues.

Index Terms— DNN, Multi-Language, Speaker Recognition

1. INTRODUCTION

During the last decade, neural networks have experienced a renaissance as a powerful machine learning tool. Deep Neural Networks (DNN) have been also successfully applied to the field of speech processing. After their great success in automatic speech recognition (ASR) [1], DNNs were also found very useful in other fields of speech processing such as speaker [2, 3, 4] or language recognition [5, 6, 7]. In speech recognition, DNNs are often directly trained for the "target" task of frame-by-frame classification of speech sounds (e.g. tied tri-phone states). Similarly, a DNN directly trained for frame-by-frame classification of languages was successfully used for language recognition in [7]. However, this system provided competitive performance only for speech utterances of short durations.

In the field of speaker recognition, DNNs are usually used in more elaborate and indirect way: One approach is to use DNNs for extracting frame-by-frame speech features. Such features are then used in the usual way (e.g. input to i-vector based system [8]). These features can be directly derived from the DNN output posterior probabilities [9] and combined with the conventional features (PLP or MFCC) [10]. More commonly, however, bottleneck (BN) DNNs are trained for a specific task, and the features are taken from a narrow hidden layer compressing the relevant information into low dimensional feature vectors [6, 5, 11]. Alternatively, standard DNN (with no bottleneck) can be used, where the high-dimensional outputs of one of the hidden layers can be converted to features using a dimensionality reduction technique such as PCA [12].

In [13], we analyzed various DNN approaches to speaker recognition (and similar studies were conducted e.g. in [14, 15]). We used two different DNN's (a mono-lingual DNN—trained on the Fisher English data corpus—and a multi-lingual DNN—trained on 11 languages of the Babel data collection). The rest of the system was trained on the PRISM set, i.e. mainly on the English data. We reported our results only on the NIST SRE 2010 telephone condition (i.e. only on English speech) via the Equal Error Rates (EERs) and the minimum DCF NIST metrics.

However, when tested on non-English test sets, we observed that the benefit of using the DNNs degraded dramatically. We used the "lan" Language Pack of the PRISM set (described later in the paper), and its Chinese subset—the "chn" pack in comparison with the originally used NIST SRE 2010 telephone condition. Not only we saw performance degradation in terms of EER and the minimum DCFs, but more so in terms of the actual DCFs, i.e. the systems produce heavily de-calibrated scores.

Our hypothesis was that when we use the DNN trained for the target language, the error rates would decrease. To match the sre10, "lan", and "chn" test conditions, we chose three DNNs, trained on: i) the Fisher English, the ii) Multilingual set, and iii) the Mandarin, respectively. However, it turned out that, apart from the Fisher English being optimal for the NIST SRE 2010 test, there was no clear correlation between the test language and the DNN training language.

This paper analyzes the problems that emerged when applying the current state-of-the-art SRE systems to non-English domains, and provides directions for future research. This work is an extension of our previous analysis, available as a technical report [16].

2. THEORETICAL BACKGROUND

2.1. i-vector Systems

The i-vectors [8] provide an elegant way of reducing large-dimensional input data to a small-dimensional feature vector while retaining most of the relevant information. The main principle is that the utterance-dependent Gaussian Mixture Model (GMM) supervector of concatenated mean vectors $\mathbf{s}$ is modeled as

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w}, \tag{1}$$

This work was supported by the DARPA RATS Program under Contract No. HR0011-15-C-0038. The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

This work was also supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0013. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

The work was also supported by Czech Ministry of Interior project No. VI20152020025 "DRAPAK" and European Union's Horizon 2020 programme under grant agreement No. 645523 BISON.

978-1-5090-4903-5/16/$31.00 ©2016 IEEE

where $\mathbf{m} = [\boldsymbol{\mu}^{(1)'}, \ldots, \boldsymbol{\mu}^{(C)'}]'$ is the Universal Background Model (UBM) GMM mean supervector (of C components), $\mathbf{T} = [\mathbf{T}^{(1)'}, \ldots, \mathbf{T}^{(C)'}]'$ is a low-rank matrix representing M bases spanning subspace with important variability in the mean supervector space, and $\mathbf{w}$ is a latent variable of size M with standard normal distribution.

The i-vector $\boldsymbol{\phi}$ is the Maximum a Posteriori (MAP) point estimate of the variable $\mathbf{w}$. It maps most of the relevant information from a variable-length observation $\mathcal{X}$ to a fixed- (small-) dimensional vector. $\mathbf{L}_{\mathcal{X}}$ is the precision of the posterior distribution.

The closed-form solution for computing the i-vector can be expressed as a function of the *zero-* and *first-order statistics*: $\mathbf{n}_{\mathcal{X}} = [N_{\mathcal{X}}^{(1)}, \ldots, N_{\mathcal{X}}^{(C)}]'$ and $\mathbf{f}_{\mathcal{X}} = [\mathbf{f}_{\mathcal{X}}^{(1)'}, \ldots, \mathbf{f}_{\mathcal{X}}^{(C)'}]'$, where

$$N_{\mathcal{X}}^{(c)} = \sum_t \gamma_t^{(c)} \tag{2}$$

$$\mathbf{f}_{\mathcal{X}}^{(c)} = \sum_t \gamma_t^{(c)} \mathbf{o}_t, \tag{3}$$

where $\gamma_t^{(c)}$ is the posterior (or occupation) probability of frame t being generated by the mixture component c. The tuple $\gamma_t = (\gamma_t^{(1)}, \ldots, \gamma_t^{(C)})$ is usually referred to as *frame alignment*. Note that this variable can be computed either using the GMM UBM or using a completely different model [2, 14, 15]. We will refer to this approach as a *DNN alignment* approach later in this paper. The i-vector is then expressed as

$$\boldsymbol{\phi}_{\mathcal{X}} = \mathbf{L}_{\mathcal{X}}^{-1}\bar{\mathbf{T}}'\bar{\mathbf{f}}_{\mathcal{X}} \tag{4}$$

where $\mathbf{L}_{\mathcal{X}}$ is the precision matrix of the posterior distribution, computed as:

$$\mathbf{L}_{\mathcal{X}} = \mathbf{I} + \sum_{c=1}^{C} N_{\mathcal{X}}^{(c)} \bar{\mathbf{T}}^{(c)'} \bar{\mathbf{T}}^{(c)}, \tag{5}$$

with c being the GMM UBM component index, and the 'bar' symbols denote normalized variables:

$$\bar{\mathbf{f}}_{\mathcal{X}}^{(c)} = \boldsymbol{\Sigma}^{(c)-\frac{1}{2}} \left(\mathbf{f}_{\mathcal{X}}^{(c)} - N_{\mathcal{X}}^{(c)} \boldsymbol{\mu}^{(c)} \right) \tag{6}$$

$$\bar{\mathbf{T}}^{(c)} = \boldsymbol{\Sigma}^{(c)-\frac{1}{2}} \mathbf{T}^{(c)}, \tag{7}$$

where $\boldsymbol{\Sigma}^{(c)-\frac{1}{2}}$ is a symmetrical decomposition (such as Cholesky decomposition) of an inverse of the GMM UBM covariance matrix $\boldsymbol{\Sigma}^{(c)}$.

2.2. Stacked Bottleneck Features (SBN)

Bottleneck Neural-Network (BN-NN) refers to such topology of a NN, one of whose hidden layers has significantly lower dimensionality than the surrounding layers. A bottleneck feature vector is generally understood as a by-product of forwarding a primary input feature vector through the BN-NN and reading off the vector of values at the bottleneck layer. We have used a cascade of two such NNs for our experiments. The output of the first network is *stacked* in time, defining context-dependent input features for the second NN, hence the term Stacked Bottleneck Features.

The NN input features are 24 log Mel-scale filter bank outputs augmented with fundamental frequency features from 4 different f_0 estimators (Kaldi, Snack[1], and two other according to [17] and [18]). Together, we have 13 f_0 related features, see [19] for more details. The conversation-side based mean subtraction is applied on the whole feature vector. 11 frames of log filter bank outputs and fundamental frequency features are stacked together. Hamming window followed by DCT consisting of 0^{th} to 5^{th} base are applied on the time trajectory of each parameter resulting in $(24 + 13) \times 6 = 222$ coefficients on the first stage NN input.

The configuration for the first NN is $222 \times D_H \times D_H \times D_{BN} \times D_H \times K$, where K is the number of targets. The dimensionality of the bottleneck layer, D_{BN} was fixed to 80. This was shown as optimal in [6]. The dimensionality of the "regular" hidden layers D_H was set to 1500. The bottleneck outputs from the first NN are sampled at times $t{-}10$, $t{-}5$, t, $t{+}5$ and $t{+}10$, where t is the index of the current frame. The resulting $80 \times 5 = 400$-dimensional features are input to the second stage NN with the same topology as first stage. The 80 bottleneck outputs from the second NN (referred as SBN) are taken as features for the conventional GMM/UBM i-vector based SID system.

We experimented with monolingual (English and Mandarin) and multilingual BN features. In the case of multilingual training, we adopted training scheme with block-softmax, which divides the output layer into parts according to individual languages. During training, only the part of the output layer is activated that corresponds to the language that the given target belongs to. See [20, 21] for detailed description.

2.3. DNN Alignment

The true frame alignment is a hidden variable in GMM modeling. Traditionally, it is computed using the GMM UBM (as used in the "baseline" and "SBN" experiments further in the paper). However, it was shown that DNNs can be used directly for posterior computation [2] .

For completeness, we report the performance of the DNN alignment systems, where the posteriors of the SBN-NNs from the previous section were used. In other words, we show the utility of the trained DNNs as both feature- and posterior-extractors.

Note that the output activation function of the Multilingual SBN is a block-softmax, giving a set of posterior probabilities (one set per training language). Therefore, we cannot utilize the Multilingual SBN for this purpose in a straightforward way.

Note also that the *normalization GMM UBM* (i.e. the $\boldsymbol{\mu}^{(c)}$ and $\boldsymbol{\Sigma}^{(c)}$ parameters) should be computed via the same DNN alignment as used in eq. (2) and (3).

3. EXPERIMENTS

3.1. DNN Training Data

For training the Multilingual neural networks, the IARPA Babel Program data[2] were mainly used. This data set simulates the scenario of what one could collect in a limited time from a completely new language. It consists mainly of conversational telephone speech (CTS), but scripted recordings, as well as far field recordings, are present. We used 11 languages to train our multilingual SBN feature extractor. The *language list* (as referred to later in this paragraph) consists of Cantonese, Assamese, Bengali, Pashtu, Turkish, Tagalog, Vietnamese, Haiti, Lao, Tamil, and Zulu. More details about the characteristics of the languages can be found in [22]. The phone-state target labels were obtained using forced-alignment with our BABEL ASR system [23], with $471 + 141 + 147 + 216 + 126 + 252 + 303 + 99 + 411 + 102 + 219 = 2487$ phone states, respectfully to the *language list*.

[1] http://kaldi.sourceforge.net, www.speech.kth.se/snack/

[2] Collected by Appen, `http://www.appenbutlerhill.com`

Table 1. *Comparison of the systems under the PRISM "lan" and "chn", and the SRE2010-condition 5 (tel-tel) tests. We expected (without result) the Multilang SBN to perform best in the "lan" condition, and a variant of Mandarin to perform best in the "chn" condition.*

Test set	System	$\mathrm{DCF}^{\mathrm{min}}_{\mathrm{new}}$		$\mathrm{DCF}^{\mathrm{min}}_{\mathrm{old}}$		EER [%]	
		male	female	male	female	male	female
chn	Baseline	0.1834	0.3019	0.0621	0.0894	1.44	2.27
	English SBN	0.1491	0.2251	**0.0418**	0.0838	**1.00**	1.99
	Mandarin SBN	0.1480	0.2368	0.0511	0.0755	1.45	2.47
	Multilang SBN	0.2121	**0.1907**	0.0439	**0.0670**	1.16	**1.93**
	English DNN	**0.1373**	0.3621	0.0616	0.1192	1.29	3.05
	Mandarin DNN	0.1688	0.2574	0.0516	0.1018	1.17	2.70
lan	Baseline	0.2979	0.9836	0.1021	**0.2007**	2.60	5.05
	English SBN	0.2963	0.9848	0.0979	0.2305	2.45	4.93
	Mandarin SBN	**0.2734**	0.9787	**0.0685**	0.2282	**1.69**	**4.11**
	Multilang SBN	0.4008	0.9854	0.0898	0.2997	2.16	5.03
	English DNN	0.2963	0.9463	0.0914	0.2228	2.70	5.68
	Mandarin DNN	0.3705	**0.9234**	0.1450	0.3255	3.57	7.14
sre10	Baseline	0.3577	0.3387	0.0967	0.1013	1.84	1.94
	English SBN	0.1295	**0.1679**	0.0387	0.0471	1.17	1.11
	Mandarin SBN	0.1459	0.2087	0.0440	0.0604	1.20	1.11
	Multilang SBN	0.1280	0.1696	0.0416	0.0544	1.21	1.16
	English DNN	**0.1200**	0.2212	**0.0352**	**0.0449**	**0.71**	**0.93**
	Mandarin DNN	0.2732	0.3356	0.0702	0.0856	1.60	1.83

For the monolingual English DNN variant, we have used a selection of 250 hours of data derived from the Fisher English Part 1 and 2 with 2423 tied tri-phone states.

For the monolingual Mandaring DNN, we have used total of 153 hours from the Mandarin HKUST, and the Mandarin CallHome/CallFriend collections [24], with 4941 tied tri-phone states.

3.2. Test Set and Evaluation Metric

We report our results on the "Language Set" pack of the PRISM set [25], referred to as "lan" later in the results. It was crafted from the NIST SRE 2005–2008 datasets by selecting 500 speakers for which there exists at least one session in a language other than English. Additional 300 speakers (that appear only in English conversations) were added from the NIST SRE 2010. The trials were created as a Cartesian product of all sessions sessions, resulting in 3590/130880 male, and 6304/297683 female target/non-target trials, respectively. Note that half of the trials are still English.

Moreover, results on the Chinese subset of the "lan" condition, referred to as "chn" are reported. The set comprises of 1027/59004 male, and 1555/113405 female target/non-target trials, respectively.

To provide a contrastive view, we also report the results on the NIST SRE 2010 data extended core condition (telephone-telephone, "condition-5"), referred to as "sre10", with 3465/175873 male, and 3704/233077 female target/non-target trials, respectively.

The detection cost function (DCF) is used as a primary evaluation metric. We report two numbers: $\mathrm{DCF}^{\mathrm{min}}_{\mathrm{old}}$and $\mathrm{DCF}^{\mathrm{min}}_{\mathrm{new}}$, corresponding to the primary evaluation metric for the NIST speaker recognition evaluation in 2008 and 2010, respectively. We also report their *actual* variants $\mathrm{DCF}^{\mathrm{act}}_{\mathrm{old}}$and $\mathrm{DCF}^{\mathrm{act}}_{\mathrm{new}}$. Equal Error Rate (EER) is also reported. For more details, see the evaluation plans of NIST SRE [3].

3.3. System Description

Voice Activity Detection (VAD) was performed using Neural Network speech/non-speech classifier. The NN was trained on Czech CTS data where we artificially added noise with different levels of SNR to 30% of the database. The NN had two hidden layers each comprising of 300 neurons. We used a vectorized block of 31 frames of 15 Mel filter bank energies as input features. For the *interview data*, we removed the interviewer based on the ASR transcripts provided by NIST.

As the baseline features, we used 19 MFCC coefficients + energy augmented with their delta and double delta coefficients, resulting in 60-dimensional feature vectors. The analysis window was 20 ms long with the shift of 10 ms. First, we removed silence frames according to VAD, after which we applied short-time (300 frames) cepstral mean and variance normalization.

The PRISM set [25] was chosen as the principal training dataset platform. It contains the following telephone data: NIST SRE 2004, 2005, 2006, 2008, 2010 Switchboard II Phases 2 and 3, Switchboard Cellular Parts 1 and 2, Fisher English Parts 1 and 2 giving 9670 female speakers. We have not included any noisy or reverberated data.

A gender-independent UBM was represented as a full or diagonal covariance 2048-component GMM. It was trained on a subset of the PRISM training set: 15602 files equally distributed between telephone and microphone conditions, and male and female portions. The variance flooring was used in each iteration of EM algorithm during the UBM training. Gender-independent i-vector extractor was trained (in 10 iterations of a joint Expectation Maximization and

[3] www.itl.nist.gov/iad/mig/tests/sre/

Table 2. *Analysis of the actual DCF's under the PRISM "lan" and "chn", and the SRE2010-condition 5 (tel-tel) tests. Note the system de-calibration on the "lan" and "chn" conditions. Also note that de-calibration is more emphasized for the female conditions. (Due to the dynamic range of the values, we prefer to report a table of numbers rather than a graph plot.)*

Test	System	$\mathrm{DCF}_{\mathrm{new}}$				$\mathrm{DCF}_{\mathrm{old}}$			
		actual		min		actual		min	
		male	female	male	female	male	female	male	female
chn	Baseline	5.7461	16.0798	0.1834	0.3019	0.1206	0.2785	0.0621	0.0894
	English SBN	1.5201	10.4024	0.1491	0.2251	0.0515	0.1857	0.0418	0.0838
	Mandarin SBN	8.4710	25.2394	0.1480	0.2368	0.1536	0.4003	0.0511	0.0755
	Multilang SBN	3.9156	12.3843	0.2121	0.1907	0.0863	0.2189	0.0439	0.0670
	English DNN	10.2419	46.4058	0.1373	0.3621	0.1856	0.6857	0.0616	0.1192
	Mandarin DNN	30.4309	75.9809	0.1688	0.2574	0.4683	0.9842	0.0516	0.1018
lan	Baseline	3.5369	14.0482	0.2979	0.9836	0.1142	0.2812	0.1021	0.2007
	English SBN	2.1503	24.4566	0.2963	0.9848	0.0702	0.3476	0.0979	0.2305
	Mandarin SBN	5.8890	30.2647	0.2734	0.9787	0.1333	0.4363	0.0685	0.2282
	Multilang SBN	5.2089	38.1320	0.4008	0.9854	0.1121	0.4855	0.0898	0.2997
	English DNN	6.6261	36.8887	0.2963	0.9463	0.1427	0.5451	0.0914	0.2228
	Mandarin DNN	16.0119	58.9831	0.3705	0.9234	0.2856	0.7746	0.1450	0.3255
sre10	Baseline	0.4323	0.3442	0.3577	0.3387	0.1587	0.2171	0.0967	0.1013
	English SBN	0.1472	0.1750	0.1295	0.1679	0.0976	0.1098	0.0387	0.0471
	Mandarin SBN	0.1815	0.2139	0.1459	0.2087	0.1264	0.1428	0.0440	0.0604
	Multilang SBN	0.1530	0.1921	0.1280	0.1696	0.1171	0.1339	0.0416	0.0544
	English DNN	0.1234	0.2286	0.1200	0.2212	0.0800	0.1204	0.0352	0.0449
	Mandarin DNN	0.3320	0.3539	0.2732	0.3356	0.1231	0.1865	0.0702	0.0856

Table 3. *The effect of calibration on the actual DCF's under the PRISM "lan" and "chn", and the SRE2010-condition 5 (tel-tel) tests for the English SBN system.*

Test	System	$\mathrm{DCF}_{\mathrm{new}}^{\mathrm{act}}$		$\mathrm{DCF}_{\mathrm{old}}^{\mathrm{act}}$	
		male	female	male	female
chn	Uncal	1.5201	10.4024	0.0515	0.1857
	Cal	0.5278	0.5080	0.0642	0.0859
lan	Uncal	2.1503	24.4566	0.0702	0.3476
	Cal	0.5519	1.2460	0.0950	0.2311
sre10	Uncal	0.1472	0.1750	0.0976	0.1098
	Cal	0.8349	0.8604	0.2087	0.2487

Minimum Divergence steps) using the entire PRISM set. The results are reported with 600-dimensional i-vectors. Gender-independent LDA and PLDA was trained on the same data as the i-vector extractor.

3.4. Results and Discussion

Tab. 1 shows the overall results of all systems in terms of (calibration insensitive) $\mathrm{DCF}_{\mathrm{old}}^{\mathrm{min}}$, $\mathrm{DCF}_{\mathrm{new}}^{\mathrm{min}}$, and EER. For the "sre10" test, the best performing system is the DNN-alignment with the DNN trained on the Fisher English data, as expected. However, when looking at the "lan" condition, there is no gain from switching from the Baseline system to English DNN (and only a negligible gain in switching to English SBN).

Our hypothesis was that this behavior would be fixed by using a more general DNN, such as the Multilingual DNN (only in the SBN variant, as explained in Sec. 2.3), since the test comprises of numerous languages. However, it turned out that Mandarin SBN suited this condition best.

Looking at the "chn" condition, we again expected the Mandarin DNN (or SBN) to significantly outperform the English and Multilang DNN's, but with no result.

Our initial hypothesis was that the English training corpus is the largest, and therefore had to provide best phone accuracy and thus a better acoustic space clustering. However, it was observed in many cases (e.g. in [26]) that better phone accuracy does not necessarily imply better SRE performance. Therefore, we leave this question open for future research.

Let us also note that the UBM/i-vector/PLDA training data are identical—i.e., mainly English—across the different systems. Our hypothesis is that even if the DNN matches the target language, the acoustic space clustering does not correspond to the observed data. Therefore, the first-order statistics (3) for the i-vector extractor computation are "warped", and the i-vector extractor captures a different "total" variability than is in fact used for the test. One of the possible indications for this hypothesis is the fact that the performance on the "sre10" condition does not vary dramatically across different systems. Similar hypothesis holds for the PLDA/LDA modeling, where the within/across variabilities are modeled using these "warped" i-vectors.

Tab. 2 shows the overall performance summary in terms of the actual vs. the minimum DCF values, i.e., it directly shows the calibration loss. We see that the "sre10" condition is well calibrated, i.e., the actual values are close enough to the minimum counterparts. However, looking at the “chn” and “lan” tests, and especially at the new DCF metric, the calibration losses are extremely high. This effect is even more pronounced for the female part of the tests.

In Tab. 3, we show the effect of a linear calibration on the English SBN system. Because of the lack of an independent held-out set, we performed a cheating (gender-independent) calibration trained using the “lan” trial set, which contains both English and Chinese trials.

We see that although not perfect, the new DCFs of the “lan” and “chn” were fixed, especially in the female case (which could be explained by having twice as many female trials compared to the male portion). It seems that even though English trials were in majority in the “lan” set, the calibration still helped the non-English trials. The “chn” calibration loss reduction was the most noticeable. The “sre10” condition got de-calibrated, as expected.

All this behavior indicates a heavy language-dependent score modality. For the time being, we do not have any solution or deeper analysis of this problem and again, we keep this issue open for future research.

4. CONCLUSIONS

In this work, we have studied the behavior of the DNN techniques in SRE i-vector/PLDA systems, currently considered to be state-of-the-art, as evaluated on the most common NIST SRE English test sets, such as the NIST SRE 2010, condition 5. We have shown that when applied to non-English test sets, these techniques stop being effective and are susceptible to de-calibration of the scores produced by the traditional i-vector/PLDA systems. We have also observed that selecting a DNN to match the test condition does not solve the issues mentioned above.

This work therefore leaves more questions than answers, and suggests that we focus on the analysis of the DNN acoustic space clustering with regard to multiple languages and other types of variability, and that we study the behavior of clustering with regard to the available SRE training data.

5. REFERENCES

[1] Geoffrey Hinton, Li Deng, Dong Yu, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath George Dahl, and Brian Kingsbury, “Deep neural networks for acoustic modeling in speech recognition,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, November 2012.

[2] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, “A novel scheme for speaker recognition using a phonetically-aware deep neural network,” in *ICASSP*, 2014.

[3] Y. Lei, L. Ferrer, M. McLaren, and N. Scheffer, “Comparative study on the use of senone-based deep neural networks for speaker recognition,” *Submitted to IEEE Trans. ASLP*, 2014.

[4] Garcia-Romero D., Zhang X., McCree A., and Povey D., “Improving speaker recognition performance in the domain adaptation challenge using deep neural networks,” in *SLT*, 2014.

[5] Y. Song et al, “i-vector representation based on bottle neck feature for language identification,” in *IEEE Electronics Letters*, 2013.

[6] Pavel Matějka et al., “Neural network bottleneck features for language identification,” in *IEEE Odyssey: The Speaker and Language Recognition Workshop*, Joensu, Finland, 2014.

[7] Ignacio Lopez-Moreno, Javier Gonzalez-Dominguez, and Oldřich Plchot, “Automatic language identification using deep neural networks,” in *ICASSP 2014*, Florence, Italy, 2014.

[8] 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.

[9] Mireia Diez, Amparo Varona, Mikel Penagarikano, Luis Javier Rodriguez-Fuentes, and Germán Bordel, “Using phone log-likelihood ratios as features for speaker recognition,” in *Interspeech 2013*, Lyon, France, 2013.

[10] Jeff Ma et al., “Improvements in language identification on the RATS noisy speech corpus,” in *Interspeech 2013*, Lyon, France, 2013.

[11] Najim Dehak Fred Richardson, Douglas A. Reynolds, “A unified deep neural network for speaker and language recognition,” in *Interspeech*, 2015.

[12] Yuan Liu, Yanmin Qian, Nanxin Chen, Tianfan Fu, Ya Zhang, and Kai Yu, “Deep feature for text-dependent speaker verification,,” *Speech Communication*, vol. 73, pp. 1–13, October 2015.

[13] Pavel Matějka, Ondřej Glembek, Ondřej Novotný, Oldřich Plchot, František Grézl, Lukáš Burget, and Jan Černocký, “Analysis of DNN approaches to speaker identification,” in *Proceedings of the 41th IEEE International Conference on Acoustics, Speech and Signal Processing*. 2016, pp. 5100–5104, IEEE Signal Processing Society.

[14] Yao Tian, Meng Cai, Liang He, and Jia Liu, “Investigation of bottleneck features and multilingual deep neural networks,” in *Interspeech*, 2015.

[15] Sandro Cumani, Oldřich Plchot, and Pietro Laface, “Comparison of hybrid DNN-GMM architectures for speaker recognition,” in *ICASSP*. 2016, IEEE Signal Processing Society.

[16] Ondřej Novotný, Pavel Matějka, Ondřej Glembek, Oldřich Plchot, František Grézl, Lukáš Burget, and Jan “Honza” Černocký, “DNN-based SRE systems in multi-language conditions,” 2016, BUT Technical Report, *http://www.fit.vutbr.cz/research/pubs/report.php?id=11235*, also being submitted to IEEE Signal Processing Letters.

[17] Kornel Laskowski and Jens Edlund, “A Snack implementation and Tcl/Tk interface to the fundamental frequency variation spectrum algorithm,” in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta, may 2010.

[18] David Talkin, “A robust algorithm for pitch tracking (RAPT),” in *Speech Coding and Synthesis*, W. B. Kleijn and K. Paliwal, Eds., New York, 1995, Elseviever.

[19] Martin Karafiát, František Grézl, Karel Veselý, Mirko Hannemann, Igor Szőke, and Jan Černocký, “BUT 2014 Babel system: Analysis of adaptation in NN based systems,” in *Interspeech 2014*, 2014, pp. 3002–3006.

[20] Karel Veselý, Martin Karafiát, František Grézl, Miloš Janda, and Ekaterina Egorova, “The language-independent bottleneck features,” in *Proceedings of IEEE 2012 Workshop on Spoken Language Technology*, 2012, pp. 336–341.

[21] Radek Fér, Pavel Matějka, Františěk Grézl, Oldřich Plchot, and Jan Černocký, “Multilingual bottleneck features for language recognition,” *Interspeech 2015*, 2015.

[22] M. Harper, “The BABEL program and low resource speech technology,” in *ASRU 2013*, Dec 2013.

[23] Martin Karafiát, František Grézl, Mirko Hannemann, Karel Veselý, and Jan ”Honza” Černocký, “BUT BABEL System for Spontaneous Cantonese,” in *Interspeech 2013*, Lyon, France, 2013, pp. 2589–2593.

[24] Martin Karafiát, Murali Karthick Baskar, František Grézl, Karel Veselý, and Jan “Honza” Černocký, “Multilingual BLSTM and SSNN adaptation in Babel,” in *submitted to SLT*, 2016.

[25] L. Ferrer, H. Bratt, L. Burget, J. 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.

[26] Alicia Lozano-Diez, Anna Silnova, Pavel Matejka, Ondrej Glembek, Oldrich Plchot, Jan Pesan, Lukas Burget, and Joaquin Gonzalez-Rodriguez, “Analysis and optimization of bottleneck features for speaker recognition,” in *Odyssey 2016: The Speaker and Language Recognition Workshop*, Bilbao, Spain, June 21-24 2016, pp. 352–357.