Scalable I-vector Concatenation for PLDA based Language Identification System

Saad Irtza[1,2], Haris Bavattichalil[1], Vidhyasaharan Sethu[1], Eliathamby Ambikairajah[1,2]

[1] School of Electrical Engineering and Telecommunications, UNSW Australia
[2] ATP Research Laboratory, National ICT Australia (NICTA), Australia
s.irtza@student.unsw.edu.au

***Abstract*— Language identification systems combining i-vectors estimated from different acoustic feature spaces have recently been shown to be superior to i-vector systems based on a single acoustic feature space. Specifically, i-vectors estimated using MFCC and PLP front-ends were concatenated prior to using LDA to obtain a combined i-vector. In this work, we investigate the scalability of this i-vector concatenation based framework to incorporate a larger number of front-ends, in particular, phonotactic front-ends. A modification to the framework is also proposed in order to improve this scalability. The proposed framework is evaluated on the 30, 10 and 3 seconds test set of NIST 2007 LRE database.**

I. INTRODUCTION

Language Identification (LID) is the task of automatically identifying the spoken language from a speech utterance. The most commonly used LID systems are those based on acoustic and phonotactic information [1-3]. Typically in systems based on acoustic information, speech signals are represented by a sequence of short term spectral or prosodic feature vectors. Longer term information is then captured by computing GMM supervector representations of the utterances. More recently, the total variability factor analysis of supervectors (the i-vector framework) has been shown to be even more effective in LID tasks [4]. The i-vector approach based on acoustic front-ends has been explored in the context of language identification with both Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) coefficients with promising results [5].

In phonotactic feature based systems, speech is tokenized into phoneme sequences which is in-turn used to derive n-gram counts as features [1, 2]. Recently, i-vectors based on Phone Log Likelihood Ratio (PLLR) features, computed using suitable phone decoders, have been utilised for LID tasks [4, 6-8]. Phone decoders based on Temporal Pattern Neural Network (TRAPs/NN) [9] are commonly used for deriving PLLR features. Among the TRAPs/NN phone decoders, Hungarian (HU), Czech (CZ) and Russian (RU) are the most commonly used ones and they extract 59, 43 and 50 dimensional PLLR features per frame, respectively.

Language identification systems that combine various types of acoustic and phonotactic features have been shown to outperform systems based on a single front-end [5-8, 10]. A straight forward approach to this combination of multiple acoustic front-ends involves the concatenation of i-vectors estimated from the different acoustic features spaces prior to classification based on the concatenated i-vector space, which was shown to be effective [10, 11]. Linear Discriminant Analysis (LDA) was then used to extract discriminative language feature vectors from the concatenated i-vectors. The i-vector concatenation framework was found to perform better than those using score level fusion of the two independent systems based on the two acoustic front-ends. Motivated by the performance of the i-vector concatenation approach based on acoustic features, this work attempts to integrate phonotactic features to that framework. Specifically, we explore the concatenation of i-vectors based on acoustic front-ends to those based on phonotactic front-ends.

A major challenge to this approach of concatenating multiple acoustic and phonotactic i-vectors is the large dimensionality of the combined i-vector. Particularly, since the traditional LDA algorithm becomes less effective when dealing with high dimensional data [12]. The '*Direct LDA*' algorithm, conceptually similar to PCA followed by LDA, was proposed to address this large dimensions small sample size (LDSS) problem in face recognition [12]. In this paper, we compare '*Direct LDA*' to the traditional LDA. Further, we propose a third approach which utilizes LDA prior to i-vector concatenation. The proposed approach is also compared with score level fusion of individual systems.

II. SYSTEM DESCRIPTION

The language identification (LID) system used in all the work reported in this paper combines i-vectors estimated from acoustic and phonotactic front-ends described below followed by a GPLDA back-end as shown in Figure 1. It should be noted that the i-vectors corresponding to each front-end are estimated using a UBM and a T-matrix (refer section II B) specific to that front-end.

A. Front-Ends

Three sets of frame based PLLR features [8] of 59, 50 and 43 dimensions are derived using the phoneme state posteriors computed using Hungarian (HU), Russian (RU) and Czech (CZ) TRAPs/NN phone recognizers [9], respectively. Following this, voice activity detection (VAD) is carried out by removing the frames whose highest PLLR value

corresponds to the non-speech unit. The PLLR features were estimated every 10ms using 25ms frames.

The acoustic front-ends used in this study are based on MFCCs and PLPs. Specifically, 13 dimensional MFCCs and 13 dimensional PLP coefficients, augmented with SDCs based on 13-7-1-3 configuration are used. The openSMILE toolbox [13] was used to for acoustic feature extraction and the corresponding voice activity detection. Both the MFCC and PLP feature vectors were estimated using hamming windows based on 20ms frames with 50% overlap.

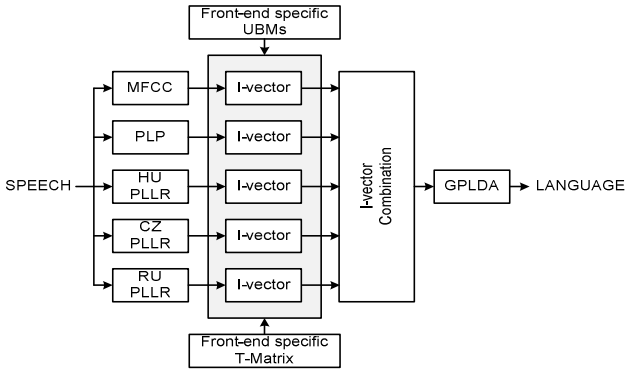

Figure 1: Overview of i-vector combination based LID system

B. *I-vectors Extraction*

Gaussian mixture models (1024 components) are utilised as Universal Background models (UBMs) corresponding to each front-end and trained using ML estimation employing binary mixture splitting. Following supervector estimation based on these UBMs, total variability matrices (T-Matrix) with 400 columns each are estimated using the procedure described in [8]. The use of column normalization of the T-matrix, which has been shown to be effective in speaker identification tasks [14], is explored for the language identification.

The T-matrix normalization technique is as described as follows:

$$\boldsymbol{T_{norm}} = \left[\frac{\boldsymbol{t_1}}{\|\boldsymbol{t_1}\|} \ldots \frac{\boldsymbol{t_R}}{\|\boldsymbol{t_R}\|}\right] \quad (1)$$

Where, $\boldsymbol{T_{norm}}$ denotes the normalized T-matrix and $\boldsymbol{t_1}, \boldsymbol{t_2}, \ldots, \boldsymbol{t_R}$ all denotes the columns of the unnormalised T-matrix. The normalized T-matrix is then used to extract i-vectors of from all training and testing utterances.

C. *LID system backend using GPLDA*

In this study, a Gaussian PLDA (GPLDA) was used as a classifier to make the final decision based on the combined i-vectors. The GPLDA back-end has shown to be effective, first in speaker identification [15] and later in LID tasks [16]. Typically, GPLDA based systems use length normalization (Figure 2) of i-vectors to overcome their non-Gaussianity [17]. The alternative T-matrix normalization (as outlined in section II B) serves a similar purpose and the two approaches are compared. In the GPLDA approach, the i-vectors are represented by a generative model given as,

$$\boldsymbol{w_u} = \bar{\boldsymbol{w}} + \boldsymbol{L x_u} + \boldsymbol{\epsilon_u} \quad (2)$$

Where, w_u denotes the i-vector corresponding to utterance u; L is the eigen-language matrix; x_u denotes the language factors corresponding to utterance u and ϵ_u denotes the utterance specific within-language variability component. In the PLDA approach, the language specific part is modeled by $\bar{w} + Lx_u$ and contains the discriminatory information for language recognition. The GPLDA parameters are estimated via maximum likelihood (ML) estimation. GPLDA scoring is performed by computing the log likelihood ratio of two hypothesis tests as follows

$$LLR = log\frac{P(w_{target}, w_{test}|H_1)}{P(w_{target} \mid H_0)P(w_{test}|H_0)} \quad (3)$$

Where, H_1 is the hypothesis that train and test i-vectors correspond to the same language while H_0 denotes the hypothesis that the train and test vectors correspond to different languages.

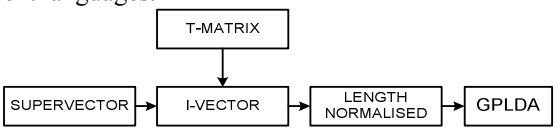

Figure 2: I-vector length normalisation

III. EXPERIMENTAL SETUP

The LID experiments reported in this work are performed on the NIST 2007 LRE dataset. The dataset consists of conversational telephonic speech in 14 languages. For training the language models and for the development purpose speech utterances derived from Call-Friend, NIST 2005 LRE and NIST 2007 LRE datasets are used. The distribution of the training, development and evaluation data used for each target language is same as that described in [8]. Total duration of the training data used is approximately 968 hours. The development test set consists of 10 conversations selected randomly from each target language. Final results are reported on 3 sec test set for the primary task in the NIST 2007 LRE dataset. The 30 , 10 and 3 sec test sets from the NISTAverage detection cost (C_{avg}) and log likelihood ratio cost (C_{llr}) as define for NIST LRE tasks are used for evaluating the performance of the LID systems.

IV. T-MATRIX NORMALIZATION VS LENGTH NORMALIZATION

In order to compare T-matrix normalization to i-vector length normalization, identical i-vector GPLDA based LID systems based on all the acoustic and phonotactic front-ends explored in this study were implemented using both normalization methods and the performances are compared in Table 1 in terms of both C_{llr} and C_{avg}. From the table, it can be seen that the T-matrix normalization is either comparable or outperforms i-vector length normalization in all cases. Therefore, T-matrix normalization is used in all following experimental work.

Table 1: Performances of LID systems based on acoustic and phonotactic front-ends using both i-vector length normalization and T-Matrix normalization.

LID Systems		I-vector Normalization			T-Matrix Normalization		
		$C_{avg}\times100/\ C_{LLR}$			$C_{avg}\times100/\ C_{LLR}$		
		30s	10s	3s	30s	10s	3s
Acoustic	MFCC (A_1)	2.76/ 0.48	6.72/ 0.75	14.70 /1.78	2.71/ 0.48	6.67/ 0.74	13.81 /1.57
	PLP (A_2)	2.89/ 0.52	6.91/ 0.82	14.64 /1.70	2.89/ 0.52	6.81/ 0.79	14.58 /1.65
Phonotactic	HU (P_1)	2.51/ 0.42	6.30/ 0.63	14.09 /1.35	2.46/ 0.41	6.22/ 0.61	13.19 /1.24
	CZ (P_2)	2.59/ 0.44	6.39/ 0.68	14.28 /1.41	2.58/ 0.44	6.37/ 0.67	13.95 /1.37
	RU (P_3)	2.55/ 0.43	6.47/ 0.70	14.56 /1.50	2.52/ 0.41	6.41/ 0.69	14.48 /1.44

V. I-VECTOR COMBINATION

A. I-vector Concatenation Prior to LDA

Figure 3 shows the block diagram of the LID system that combines i-vectors from all five front-ends by concatenating the individual i-vectors prior to dimensionality reduction using LDA. Specifically, 400 dimensional i-vectors corresponding to each of the five front-ends are concatenated resulting in a 2000 dimensional vector. The subsequent LDA reduces this dimensionality to 1200. The resultant combined i-vectors are then used by the GPLDA back-end for language identification.

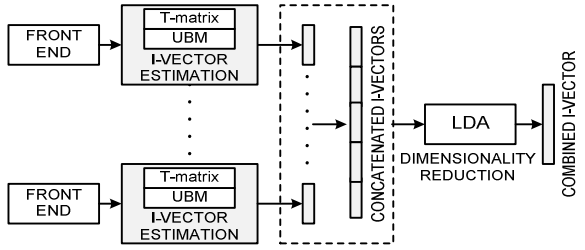

Figure 3: Block diagram of concatenation + LDA scheme for i-vector combination

Table 2 shows the performance of the LID system based on concatenated i-vectors for different combinations of front-ends.

Table 2: Performances of the concatenation + LDA based LID systems

Fusion of LID systems	I-vector concatenation before LDA		
	$C_{avg}\times100/\ C_{LLR}$		
	30s	10s	3s
(P_1)+(P_2)+(P_3)	4.3/0.59	9.2/0.91	14.78/1.89
(P_1)+(P_2)+(P_3)+(A_1)	6.4/0.71	13.5/1.1	17.59/2.07
(P_1)+(P_2)+(P_3)+(A_1)+(A_2)	7.1/0.83	14.8/1.27	19.12/2.81

In Table 2, the first column describes the fusion of individual systems listed in Table 1. It can be seen that the performance of the overall system degrades as more front-ends are integrated into the system. This is counter-intuitive since more information is available to the back-end as more front-ends are integrated and suggests that the concatenation followed by LDA approach to the combination of i-vectors is not scalable enough to accommodate multiple acoustic and phonotactic front ends. A major challenge in this approach is the high dimensionality of the combined i-vectors, since the traditional LDA algorithm becomes less effective when dealing with high dimensional data.

B. Direct LDA

In order to address the shortcomings of LDA when dealing with high dimensional data, Principle Component Analysis (PCA) has been used to reduce the dimensionality prior to LDA. However, a potential problem with this two stage, PCA+LDA approach is that the PCA step may remove discriminative information that may be useful for the subsequent LDA step. Direct LDA was proposed as an alternative one step approach in [12]. It should be noted that in an N-class problem, direct LDA provides an $N-1$ dimensional projection. In the experiments reported in this paper, there are 14 target languages and consequently the direct LDA reduces the 2000 dimensional concatenated i-vector to a 13 dimensional combined i-vector.

Direct LDA reverses the order of diagonalization of the within class scatter matrix (S_w) and the between class scatter matrix (S_b). The key idea behind this approach is to discard the null space of S_b, which contains minimal useful information, instead of the null space of S_w. Consequently, the direct LDA is similar to the PCA+LDA approach, but the first step becomes equivalent to PCA obtained when replacing S_b with total scatter matrix ($S_T = S_b + S_w$).

Given the potential advantages of direct LDA over the traditional LDA when dealing with high dimensional data, the use of direct LDA in the i-vector combination scheme outlined in Figure 3 for dimensionality reduction was evaluated. Specifically, the LDA was replaced by direct LDA and the two systems are compared.

Table 3 shows the performance of the system using direct LDA. It can be seen that when direct LDA is used for dimensionality reduction, the performance of the system increases as more front-ends are incorporated. The results show 70.7% (30s), 59.7% (10s) and 40.2% (3s) relative improvement when using direct LDA over traditional LDA on the five individual combined i-vectors systems reported in Tables 2 and 3.

Table 3: Performances of the concatenation + direct LDA based LID systems

Fusion of LID systems	Direct LDA		
	$C_{avg}\times100/\ C_{LLR}$		
	30s	10s	3s
(P1)+(P2)+(P3)	2.33/0.37	6.15/0.61	12.42/1.19
(P1)+(P2)+(P3)+(A1)	2.19/0.31	6.02/0.58	12.03/1.12
(P1)+(P2)+(P3)+(A1)+(A2)	2.08/0.28	5.96/0.55	11.39/1.02

C. Proposed I-vector Concatenation after LDA

Here, we propose an alternate method to concatenate the i-vectors derived from acoustic and phonotactic features. In the proposed framework, LDA is applied on the i-vectors prior to concatenation as shown in Figure 4.

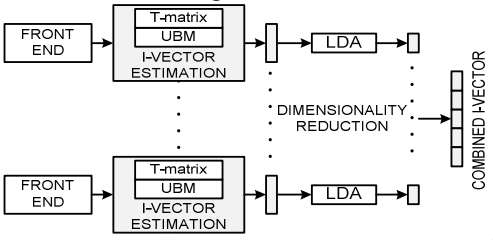

Figure 4: Block diagram of the proposed LDA + concatenation scheme for i-vector combination

Specifically, the traditional LDA is used to reduce the dimensionality of the i-vectors estimated from each front-end initially before concatenating them to form a low dimensional combined i-vector. The combined i-vectors are then input to a GPLDA back end for language identification. In the experiments reported here, the i-vector dimensionality for each front-end was reduced from 400 to 60.

Table 4 lists the performance of the LID system based on the proposed scheme of front-end specific i-vector dimensionality reduction via LDA prior to concatenation. Comparing Tables 2, 3 and 4, it can be seen that the proposed approach outperforms both the concatenation + LDA (traditional) approach as well as the concatenation + direct LDA approach for combining i-vectors from multiple acoustic and phonotactic front-ends. The proposed system achieves a relative improvement of 50.04% (30s), 19.6% (10s) and 11.7% (3s) over the concatenation + direct LDA approach for the five individual combined i-vector systems reported in Tables 3 and 4.

Table 4: Performances of the LDA + concatenation based LID systems

Fusion of LID systems	I-vector combination after LDA		
	$C_{avg}\times100/C_{LLR}$		
	30s	10s	3s
(P1)+(P2)+(P3)	1.92/0.25	5.94/0.54	11.88/1.06
(P1)+(P2)+(P3)+(A1)	1.58/0.21	5.37/0.49	10.97/0.98
(P1)+(P2)+(P3)+(A1)+(A2)	1.03/0.16	4.76/0.42	10.05/0.92

Finally, Table 5 shows the performance of score level fusion of individual systems acoustic and phonotactic systems. It can be seen that the proposed system outperforms the score level fusion of five systems by 36% (30s), 7.7% (10s) and 7.6% (3s).

Table 5: Performances of the LDA using score level fusion

Fusion of LID systems	Score level fusion		
	$C_{avg}\times100/C_{LLR}$		
	30s	10s	3s
(P1)+(P2)+(P3)	2.11/0.30	6.04/0.59	12.19/1.17
(P1)+(P2)+(P3)+(A1)	1.89/0.23	5.77/0.50	11.57/1.09
(P1)+(P2)+(P3)+(A1)+(A2)	1.61/0.21	5.16/0.47	10.88/0.95

VI. CONCLUSIONS

This paper addresses the combination of i-vectors from multiple acoustic and phonotactic front-ends for language identification. Specifically, it identifies that the established approach of concatenating i-vectors prior to dimensionality reduction via LDA is not suitable when the number of front-ends increase so as to make the dimensionality of the concatenated i-vectors very high. The paper then shows that a dimensionality reduction method explicitly designed for high dimensional data such as direct LDA is more suited. Finally, an alternative approach is proposed whereby dimensionality reduction is carried out for the i-vectors from each front-end independently prior to concatenation. Experimental results show that this proposed alternative approach outperforms both the established approach and the use of direct LDA. Specifically, in the case of the system that combines i-vectors from three phonotactic and two acoustic front-end, the proposed approach outperforms the score level fusion.

REFERENCES

[1] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language identification: A tutorial," *Circuits and Systems Magazine, IEEE,* vol. 11, pp. 82-108, 2011.

[2] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: from fundamentals to practice," *Proceedings of the IEEE,* vol. 101, pp. 1136-1159, 2013.

[3] D. A. Reynolds, W. M. Campbell, W. Shen, and E. Singer, "Automatic language recognition via spectral and token based approaches," in *Springer Handbook of Speech Processing*, ed: Springer, 2008, pp. 811-824.

[4] M. Soufifar, M. Kockmann, L. Burget, O. Plchot, O. Glembek, and T. Svendsen, "iVector Approach to Phonotactic Language Recognition," in *INTERSPEECH*, 2011, pp. 2913-2916.

[5] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak, "Language Recognition via i-vectors and Dimensionality Reduction," in *INTERSPEECH*, 2011, pp. 857-860.

[6] L. F. D'Haro, R. Cordoba, C. Salamea, and J. Ferreiros, "Language Recognition using Phonotactic-based Shifted Delta Coefficients and Multiple Phone Recognizers," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[7] L. D'Haro, R. Cordoba, C. Salamea, and J. Echeverry, "Extended phone log-likelihood ratio features and acoustic-based i-vectors for language recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 5342-5346.

[8] M. Díez, A. Varona, M. Peñagarikano, L. J. Rodríguez-Fuentes, and G. Bordel, "On the use of phone log-likelihood ratios as features in spoken language recognition," in *SLT*, 2012, pp. 274-279.

[9] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, Faculty of Information Technology, Brno University of Technology, http://www.fit.vutbr.cz/, Brno, Czech Republic, 2008.

[10] Z.-Y. Li, W.-Q. Zhang, L. He, and J. Liu, "Complementary combination in i-vector level for language recognition," in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

[11] Z.-Y. Li, W.-Q. Zhang, and J. Liu, "Multi-resolution time frequency feature and complementary combination for short utterance speaker recognition," *Multimedia Tools and Applications,* vol. 74, pp. 937-953, 2015.

[12] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data—with application to face recognition," *Pattern recognition,* vol. 34, pp. 2067-2070, 2001.

[13] Florian Eyben, Martin Wöllmer, Björn Schuller: "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", In Proc. ACM Multimedia (MM), ACM, Florence, Italy, ACM, ISBN 978-1-60558-933-6, pp. 1459-1462, October 2010. doi:10.1145/1873951.1874246

[14] W. Rao, M. W. Mak, and K. A. Lee, "Normalization of total variability matrix for i-vector/plda speaker verification," ICASSP, 2015.

[15] A. Kanagasundaram, D. Dean, J. Gonzalez-Dominguez, S. Sridharan, D. Ramos, and J. Gonzalez-Rodriguez, "Improving the PLDA based speaker verification in limited microphone data conditions," in *In Proceedings of the 14th Annual Conference of the International Speech Communication Association*, 2013, pp. 3674-3678.

[16] S. Irtza, V. Sethu, P. Le, E. Ambikairajah and H. Li " Phonemes Frequency based PLLR Dimensionality Reduction for Language Recognition," accepted in *INTERSPEECH*, 2015.

[17] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems," in *INTERSPEECH*, 2011, pp. 249-252.