i-vector Based Speaker Recognition on Short Utterances

Ahilan Kanagasundaram, Robbie Vogt, David Dean, Sridha Sridharan, Michael Mason

Speech and Audio Research Laboratory
Queensland University of Technology, Brisbane, Australia
{a.kanagasundaram, r.vogt, d.dean, s.sridharan, m.mason}@qut.edu.au

Abstract

Robust speaker verification on short utterances remains a key consideration when deploying automatic speaker recognition, as many real world applications often have access to only limited duration speech data. This paper explores how the recent technologies focused around total variability modeling behave when training and testing utterance lengths are reduced. Results are presented which provide a comparison of Joint Factor Analysis (JFA) and i-vector based systems including various compensation techniques; Within-Class Covariance Normalization (WCCN), LDA, Scatter Difference Nuisance Attribute Projection (SDNAP) and Gaussian Probabilistic Linear Discriminant Analysis (GPLDA). Speaker verification performance for utterances with as little as 2 sec of data taken from the NIST Speaker Recognition Evaluations are presented to provide a clearer picture of the current performance characteristics of these techniques in short utterance conditions.

Index Terms: speaker verification, short utterance, i-vector, SDNAP, WCCN, LDA, Gaussian PLDA

1. Introduction

The significant amount of speech required for speaker model enrolment and verification, especially in the presence of large intersession variability, has limited the widespread use of speaker verification technology in everyday applications. Continuous research on this field has been ongoing to address the robustness of speaker verification technologies under such conditions. Reducing the amount of speech required while obtaining the satisfactory performance has been the focus in a number of recent studies focused on Joint Factor Analysis (JFA) and SVM based speaker verification. These studies have shown that while performance degrades considerably in very short utterances ($< 10s$) for both approaches, JFA [1] appears to a better choice in these conditions than SVMs [2]. This paper will focus on whether a recently proposed factor-analysis front-end approach to speaker verification, called *i-vectors* [3], could form a suitable foundation for continuing research into short utterance speaker verification.

JFA originally proposed by Kenny [4], has recently evolved as a powerful tool in speaker verification to model the inter-speaker variability and to compensate for channel/session variability in the context of high-dimensional Gaussian Mixture Model (GMM) supervectors. More recently a new front-end factor analysis technique, termed i-vector (for intermediate-size vector) extraction, proposed by Dehak *et al.*[5], has evolved from JFA. Rather than taking the JFA approach of modelling a speaker and channel variability space, the i-vector approach forms a low-dimensional total-variability space that models both speaker and channel variability. The i-vector approach proposed in [3] also has the advantage that scoring uses a simple Cosine Similarity Scoring (CSS) kernel directly to perform verification, making the scoring process faster and less complex than other speaker verification methods, including JFA or Support Vector Machines (SVM) supervector approaches. Because the total variability space does contain channel variability information, i-vector speaker verification systems need to be combined with intersession compensation techniques such as Within-class Covariance Normalisation (WCCN), Linear Discriminant Analysis (LDA) and Nuisance Attribute Projection (NAP) [3]. More recently, Kenny *et al.* [6] have developed a new technique called Gaussian Probabilistic LDA (GPLDA), which divides the i-vector space into speaker and session variability subspaces, which has shown significant promise for intersession compensation for i-vector speaker verification.

The main aim of this paper is to investigate the effect that short duration utterances have on both enrolment and training when using the i-vector approach. As this approach is based on defining only one variability space, instead of the separate channel and speaker spaces of the JFA approach, we expect that i-vectors won't loose any speaker information with reduction of utterance duration [7]. We also report on the short utterance duration performance when various intersession variability compensation techniques such as WCCN, LDA and NAP are used in conjunction with i-vectors, including a short investigation of scatter-difference NAP (SDNAP), which has not yet been considered for i-vector compensation. In addition we compare the above combination of techniques with GPLDA [6]. The experimental results presented give useful indication of how the systems degrade as training and testing utterance lengths are reduced.

Section 2 of the paper provides a condensed introduction to each of the techniques investigated, from JFA, through the various front end factor analysis techniques used with total variability modeling and finally GPLDA. Section 3 provides details of the experimental procedures used to investigate the short utterance performance and presents the outcomes of the three experiments for discussion.

2. Factor analysis speaker verification

Factor analysis approaches to speaker verification were originally intended to model the intersession variability directly in the construction of the super-vectors used for scoring verification trials, such as in the standard JFA approach [4]. More recently factor analysis techniques have been considered as a front-end to form a low-dimensional total-variability subspace that can be classified more efficiently than the high-dimensional supervectors used in JFA and SVM speaker verification systems. This section will briefly outline the JFA approach, followed by a more detailed description of the common approaches

28 – 31 August 2011, Florence, Italy

to front-end factor analysis in the i-vector and GPLDA approaches.

2.1. Joint factor analysis

The factor analysis technique proposed by Kenny [4] is based on the decomposition of a speaker-dependent GMM super-vector, $\boldsymbol{\mu}$, into separate speaker and channel dependent parts (**S** and **C** respectively):

$$\boldsymbol{\mu} = \mathbf{S} + \mathbf{C}. \quad (1)$$

The speaker dependent and channel dependent components can then be represented by

$$\mathbf{S} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{D}\mathbf{z}, \quad (2)$$

$$\mathbf{C} = \mathbf{U}\mathbf{x}. \quad (3)$$

In the speaker dependent component, **m** is a session and speaker independent supervector (extracted from a universal background model (UBM) trained on a large development set), **V** is a low rank matrix representing the primary directions of speaker variability, or *eigenvoices*, and **D** is a diagonal matrix modelling the residual variability not captured by the speaker subspace. The speaker factors, **y**, and speaker residuals, **z**, are both independent random vectors having standard normal distributions. Similarly, the channel dependent component contains a low rank matrix, **U**, representing the primary directions of channel variance, or *eigenchannels*, multiplied by the channel factor vector **x**, a normally distributed random vector.

JFA speaker enrolment is performed by calculating the full speaker-dependent GMM supervectors and discarding the channel dependent component. During verification, the channel-dependent component can be estimated directly from the testing utterances, and the entire supervector can be efficiently scored using the linear dot-product approach pioneered by Glembek *et al.* [8].

2.2. i-vectors

Inspired by the earlier use of JFA speaker factors directly as features for SVM classification, Dehak *et al.* [7] have recently proposed a new approach to front-end factor analysis, termed i-vectors. Unlike the separate speaker and channel dependent subspaces of JFA, i-vectors represent the GMM super-vector by a single *total-variability* space. This single-subspace approach was motivated by the discovery in Dehak *et al.* [7] that the channel space of JFA contains information that can be used to distinguish between speakers. An i-vector speaker and channel dependent GMM super-vector can be represented by

$$\boldsymbol{\mu} = \mathbf{m} + \mathbf{T}\mathbf{w}, \quad (4)$$

where **m** is the same UBM supervector used in the JFA approach and **T** is a low rank matrix representing the primary directions of variability across all development data. The total-variability factors, **w**, is a independent normally-distributed random vector.

While i-vectors were originally considered as a feature for SVM classification, fast scoring approaches using a cosine kernel directly as a classifier, in order to produce a cosine similarity score (CSS), were found to provide similar performance to SVMs with a considerable increase in efficiency [7]. The CSS operates by comparing the angles between a test i-vector, w_{test}, and a target i-vector w_{target}:

$$\text{score}(w_{\text{target}}, w_{\text{test}}) = \frac{\langle w_{\text{target}}, w_{\text{test}} \rangle}{\|w_{\text{target}}\| \, \|w_{\text{test}}\|}. \quad (5)$$

As the total variability space represented by **T** contains both speaker and channel variability, i-vector approaches require additional intersession compensation approaches before scoring to attenuate the effects of channel variability. A number of existing approaches borrowed from SVM speaker verification such as WCCN, LDA and NAP have shown promise for this task, and will be outlined in the remainder of this section.

2.2.1. WCCN

WCCN is used as a channel compensation technique to scale a subspace in order to attenuate dimensions of high within-class variance. For use in speaker verification, a within-class covariance matrix, $\mathbf{S}_w$, is calculated using

$$\mathbf{S}_w = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (\mathbf{w}_i^s - \bar{\mathbf{w}}_s)(\mathbf{w}_i^s - \bar{\mathbf{w}}_s)^T, \quad (6)$$

where $\bar{\mathbf{w}}_s$ is the mean i-vector for each speaker, S is the total number of speakers and n_s is number of utterances of speaker s. In evaluation, the inverse of $\boldsymbol{S}_w$ is used to normalize the direction of the projected i-vector components, which is equivalent to scaling the subspace by the matrix $\boldsymbol{B}$, where $\boldsymbol{B}\boldsymbol{B}^T = \boldsymbol{S}_w^{-1}$.

2.2.2. LDA + WCCN

One of the disadvantages of the WCCN approach is that while it focuses on attenuating dimensions of high within-class variability, it can also remove information about the between-class variability that is also contained within the attenuated dimensions. In order to attempt to alleviate this problem with the WCCN approach, a transformation matrix can be trained using LDA to transform the i-vectors into a new subspace that seeks to both minimises the within class variance, defined in (6), and maximises the between class variance, $\mathbf{S}_w$, defined as

$$\mathbf{S}_b = \sum_{s=1}^{S} (\bar{\mathbf{w}}_s - \bar{\mathbf{w}})(\bar{\mathbf{w}}_s - \bar{\mathbf{w}})^T, \quad (7)$$

where $\bar{\mathbf{w}}$ is the mean over all training i-vectors.

Once the LDA transformation matrix, **A**, has been calculated the WCCN transformation can then calculated in the reduced subspace taken from $\mathbf{A}^T\mathbf{w}$, and similarly in evaluation.

2.2.3. SDNAP + WCCN

NAP can be used to combat the session variation in a similar manner to WCCN. However, rather than weighting the i-vector dimensions, NAP attempts to remove the unwanted within-class variations from the feature vector. A NAP transformation matrix traditionally has the form

$$\mathbf{P} = \boldsymbol{I} - \mathbf{V}\mathbf{V}^T \quad (8)$$

where **I** is the identity matrix and the projection matrix **V** can be obtained by taking the top N values from an eigen decomposition of the within class covariance matrix, $\mathbf{S}_w$, defined in (6).

Similar to the WCCN approach, NAP has a disadvantage that the unwanted variations due to within-class variance can have a side-effect of removing useful between class variance. A discriminative form of NAP called scatter-difference NAP (SDNAP) has recently been developed that attempts to trade-off the between and within class variance [9]. While this approach has shown promise for SVM speaker verification, it has not yet been studied in the i-vector approach.

While the calculation of **P** remains the same in SDNAP, **V** is obtained by an eigen-decomposition of a combined covariance matrix,

$$\mathbf{S} = \mathbf{S}_w - m\mathbf{S}_b, \tag{9}$$

where m is a parameter defining the relative influence of the within and between class covariance matrices.

For the system outlined later in this paper, the SDNAP projection operates has been found to work best in cooperation with WCCN. Similarly to the LDA + WCCN approach outlined previously, the i-vectors are first projected into a reduced dimensionality NAP space[1], followed by calculation of the WCCN transformation for use in evaluation.

2.3. GPLDA

Rather than attempting to compensate for intersession variability, a more sophisticated attempt to directly model session and speaker variability within the i-vector space was recently proposed by Kenny [6] as PLDA. This approach can be seen to be very similar to the JFA approach, but using i-vectors rather than GMM supervectors as the basis for factor modelling. Similarly to the JFA equations outlined in Section 2.1, a speaker and channel dependent i-vector, **w** can be defined as

$$\mathbf{w} \quad = \quad \bar{\mathbf{w}} + \mathbf{U}_1\mathbf{x}_1 + \mathbf{U}_2\mathbf{x}_2 + \boldsymbol{\varepsilon} \tag{10}$$

where U_1 is the eigenvoice matrix and U_2 is the eigenchannel matrix. $\mathbf{x}_1$ and $\mathbf{x}_2$ are the speaker and channel factors respectively and ε is the speaker residuals. Kenny investigated using both standard normal and heavy-tailed distributions for $\mathbf{x}_1$, $\mathbf{x}_2$, and ε, but we will only be investigating the Gaussian case for this paper.

GPLDA based i-vector system scoring calculated using batch likelihood ratio [6]. Batch likelihood calculation is computationally more expensive than CSS. Given two i-vectors w_{target} and w_{test}, batch likelihood ratio can be calculated as follows,

$$\ln \frac{P(w_{target}, w_{test} \mid H_1)}{P(w_{target} \mid H_0)P(w_{test} \mid H_0)} \tag{11}$$

where H_1: The speakers are same, H_0: The speaker are different

3. Analysis of short utterance performance

3.1. Methodology

13 feature-warped MFCC with appended delta coefficients and two gender dependent universal background models (UBM) containing 512 Gaussians are used throughout our experiments. These UBM were trained on NIST 2004 Speaker Recognition Evaluation (SRE) corpus. Speaker and session variability subspaces of dimension $R_y = 400$ and $R_x = 100$ are applied for JFA experiments. Total variability subspace of dimension $R_w = 400$ is applied for i-vector experiments. These total variability space, channel compensation techniques such as WCCN, LDA, NAP, SDNAP are trained on NIST 2004 Speaker Recognition Evaluation (SRE), NIST 2005 SRE and Switch board II. Speaker variability subspace of dimension of $R_{U1} = 300$ is used for GPLDA based experiments. From the experiments, it has been found that best value of SDNAP control parameter (m) is equal to 0.375.

[1]That is, $\mathbf{P} = \mathbf{V}$, where **V** is taken from the remaining eigenvectors of **S** after the top N are removed, as opposed to the full-space projection defined in (8)

Table 1: *Comparison of JFA and i-vector systems on the common set of the 2008 NIST SRE short2-short3, short2-10sec and 10sec-10sec conditions.*

System	**short2-short3**		**short2-10sec**		**10sec-10sec**	
	EER	DCF	EER	DCF	EER	DCF
JFA	3.37%	**0.0149**	9.16%	0.0390	16.69%	**0.0686**
WCCN only	3.95%	0.0189	8.86%	0.0439	17.86%	0.0705
LDA-WCCN	3.54%	0.0179	8.99%	0.0416	17.39%	0.0694
NAP-WCCN	3.71%	0.0177	8.71%	0.0415	17.71%	0.0794
SDNAP-WCCN	3.62%	0.0166	8.55%	0.0418	17.42%	0.0698
GPLDA	**3.13%**	0.0168	**7.57%**	**0.0389**	**16.40%**	0.0705

NIST 2008 SRE telephone based utterances from the short2-short3, short2-10sec and 10sec-10sec conditions were used for experiments. The truncated utterances were obtained by truncating the utterances of the NIST2008 short2-short3 condition to the specified length of active speech data for both training and testing. For NIST 2008 SRE, det condition 7 [10] was evaluated corresponding tel-tel trials.

3.2. Experiments

Initial experiments were carried out to compare the performance of channel compensation based i-vector systems, a GPLDA based i-vector system and a JFA system with standard NIST conditions. A second set of experiments were carried to compare the performance of the LDA + WCCN, SDNAP + WCCN, GPLDA and JFA systems with truncated training and testing utterances and full training and truncated testing utterances. ZT normalization was applied to channel compensation based i-vector system and JFA system, whereas S normalization was applied to the GPLDA based i-vector system, as defined in [6].

3.3. Results and Discussions

Table 1 presents results comparing the JFA system, the four alternative channel compensation based i-vector systems, and the GPLDA system on the standard NIST SRE 08 evaluation condition.

Comparing the different systems generally, the results indicate that opting for the more computationally efficient scoring processes of CSS with a total variability based system may result in a marginal drop in performance. Whether the efficiency gains outweigh the potential performance difference would be an application dependant design decision. Comparing the performance of SDNAP + WCCN to NAP + WCCN specifically, the results indicate the gains found for SDNAP over NAP in SVMs [9] hold in the i-vector systems as well. Of the systems using log-likelihood scoring, JFA & GPLDA provide similar performance on the matched training and testing conditions, while GPLD appears to be potentially more robust than JFA to variations in training and testing lengths (short2-10s).

Tables 2(a) and 2(b) presents the results comparing JFA to channel compensation based i-vector systems, LDA + WCCN and SDNAP + WCCN, and GPLDA for the truncated training,testing and full training, truncated testing conditions respectively. Overall, the results show that as the utterance length decreases, performance degrades at an increasing rate, rather than

Table 2: *Comparison of JFA and i-vector systems on the common subset of the 2008 NIST SRE short2-short3 condition with (a) truncated training and testing and (b) truncated testing only. The best performing systems by both EER and DCF are highlighted across each row.*

(a) *truncated training and testing*

Utterance Length	**JFA System**		**LDA + WCCN**		**SDNAP + WCCN**		**GPLDA System**	
(training-testing)	EER	DCF	EER	DCF	EER	DCF	EER	DCF
2 sec - 2 sec	**35.25%**	0.0988	35.35%	**0.0986**	35.67%	0.0999	36.16%	0.0999
4 sec - 4sec	30.48%	**0.0934**	31.05%	0.0966	**30.23%**	0.0968	31.30%	0.0991
8 sec - 8 sec	**23.39%**	0.0803	23.95%	**0.0800**	23.56%	0.0801	23.56%	0.0837
10 sec - 10sec	21.17%	**0.0738**	21.56%	0.0741	20.84%	0.0737	**20.34%**	0.0762
20 sec - 20sec	12.79%	0.0533	13.41%	0.0530	12.84%	**0.0528**	**11.87%**	0.0532
50 sec - 50 sec	6.51%	**0.0266**	6.44%	0.0310	6.42%	0.0299	**5.77%**	0.0272
full (2.5min) - full	3.37%	**0.0149**	3.54%	0.0179	3.62%	0.0166	**3.13%**	0.0168

(b) *full training and truncated testing*

Utterance Length	**JFA System**		**LDA + WCCN**		**SDNAP + WCCN**		**GPLDA System**	
(training-testing)	EER	DCF	EER	DCF	EER	DCF	EER	DCF
full - 2 sec	22.48%	**0.0773**	22.01%	0.0783	**21.98%**	0.0792	22.66%	0.0835
full - 4 sec	17.96%	**0.0633**	17.38%	0.0662	17.46%	0.0662	17.38%	0.0695
full - 8 sec	13.43%	**0.0492**	13.59%	0.0493	13.51%	0.0494	**12.36%**	0.0508
full - 10 sec	12.11%	0.0455	12.19%	**0.0451**	12.11%	0.0456	**11.20%**	0.0455
full - 20 sec	7.67%	0.0321	8.22%	0.0338	8.40%	0.0324	**7.34%**	**0.0313**
full - 50 sec	4.54%	**0.0200**	4.94%	0.0241	4.94%	0.0228	**4.14%**	0.0209
full - full	3.37%	**0.0149**	3.54%	0.0179	3.62%	0.0166	**3.13%**	0.0168

in proportion with the reduced length. No single technique appears to provide more resilience to this effect than any other, and though the difference in scores between the systems may appear to narrow at very short utterance lengths ($< 10s$) it is difficult to conclude this difference is significant without further exploration.

4. Conclusion

The challenges of providing robust speaker verification for applications with access to only short speech utterances remains a key hurdle to the broad adoption of speech verification systems. This paper has presented a study investigating how the current selection of factor analysis techniques perform when utterance lengths are significantly reduced. Overall, the current factor analysis approaches have not provided any clear differences in performance for short speech, with the alterative between log-likelihood based JFA and GPLDA offering marginally better performance to LDA + WCCN or SDNAP + WCCN based i-vector systems in lieu of the efficiencies available through operating in the lower-dimensional i-vector space. All the systems still exhibit performance which declines sharply once utterance lengths fall below 10s. Problems of very short utterance with factor analysis approaches will be investigated in future.

5. Acknowledgements

This project was supported by the Cooperative Research Centre for Advanced Automotive Technologies (AutoCRC).

6. References

[1] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *Interspeech 2008*, pp. 853–856, September 2008.

[2] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, "Experiments in SVM based speaker verification using short utterances," *Odyssey Speaker and Language Recognition Workshop*, 2010.

[3] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques," *Odyssey Speaker and Language Recognition Workshop*, 2010.

[4] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[5] S. Shum, N. Dehak, R. Dehak, and J. Glass, "Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification," *Submitted to ODYSSEY*, 2010.

[6] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proceedings of the Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic*, 2010.

[7] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proceedings of Interspeech*, pp. 4237 – 4240, 2009.

[8] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 4057 –4060, April 2009.

[9] R. Vogt, S. Kajarekar, and S. Sridharan, "Discriminant NAP for SVM speaker recognition," in *Proc. of Odyssey*, pp. 629 – 632, Citeseer, 2008.

[10] "The NIST year 2008 speaker recognition evaluation plan," tech. rep., NIST, April 2008.