Unimodal late fusion for NIST *i*-vector challenge on speaker detection

Hazrat Ali, Artur S. d'Avila Garcez, Son N. Tran, Xianwei Zhou and Khalid Iqbal

Speaker detection is a very interesting machine learning task for which the latest *i*-vector challenge has been coordinated by the National Institute of Standards and Technology (NIST). A simple late fusion approach for the speaker detection task on the *i*-vector challenge is presented. The approach is based on the late fusion of scores from the cosine distance method (the baseline) and the scores obtained from linear discriminant analysis. The results show that by adapting the simple late fusion approach, the framework can outperform the baseline score for the decision cost function on the NIST *i*-vector machine learning challenge.

Introduction: Researchers continue to strive for improvement of speaker evaluation systems. The latest *i*-vector challenge organised by the National Institute of Standards and Technology (NIST) [1] provided an excellent platform for researchers to submit systems with an improvement in the speaker recognition performance. The NIST *i*-vector challenge coordinated by the NIST is a continuation of the NIST speaker recognition evaluations. The challenge is based on *i*-vectors [2] which are considered to be state-of-the-art features for speaker recognition systems. The *i*-vector (each *i*-vector in the challenge is a 600 dimension vector) is a name assigned to the representation of speech utterance using total factors. The challenge dataset provides a set of target speakers where each target speaker is modelled by five *i*-vectors [1]. The speakers in the test set are defined by single *i*-vectors. The target speakers set consists of 1306 speakers and the test data consists of 9634 *i*-vectors. Each trial is composed of a target speaker model and a test *i*-vector, thus leading to a total of 12 582 004 trials. The decision algorithm has to decide whether a particular test speaker exists in the target data and reflect the level of confidence in its belief.

The baseline system for the challenge is based on a variant of cosine scoring, where *i*-vectors are projected into a unit sphere. The five *i*-vectors for each training example are averaged to obtain a mean *i*-vector. Following this, the dot product of the averaged *i*-vectors and the test data *i*-vectors provide the corresponding score for each trial. The cosine distance scoring uses the value of the cosine kernel between the target speaker *i*-vector $\boldsymbol{w}_{\text{target}}$ and the test *i*-vector $\boldsymbol{w}_{\text{test}}$ as a decision score [2, 3].

A cosine kernel can be defined by the equation

$$k(w_1,\ w_2) = \frac{\langle w_1,\ w_2 \rangle}{\|w_1\|\|w_2\|} \tag{1}$$

where w_1 and w_2 are the two *i*-vectors. Following this definition, the cosine distance scoring between the target *i*-vector $\boldsymbol{w}_{\text{target}}$ and the test *i*-vector $\boldsymbol{w}_{\text{test}}$ is given below

$$\text{score1}\left(\boldsymbol{w}_{\text{target}},\ \boldsymbol{w}_{\text{test}}\right) = \frac{\langle \boldsymbol{w}_{\text{target}},\ \boldsymbol{w}_{\text{test}} \rangle}{\|\boldsymbol{w}_{\text{target}}\|\|\boldsymbol{w}_{\text{test}}\|} \tag{2}$$

Method: Linear discriminant analysis (LDA) [4] is considered to be a simple and robust method for classification. It maximises the ratio of inter-class variance to the intra-class variance to achieve maximum separability. The LDA transforms the data into the new space and then the Euclidean distance is calculated over each test example. Generally, for n classes, n Euclidean distances are calculated for each test example. The smallest distance then determines the class for the test example. For the speaker recognition task, discriminant analysis maximises the variance between the speakers, which is a key for speaker identification.

Performance metric: The performance metric for the NIST *i*-vector challenge is the decision cosine function (DCF), which is defined by the following expression:

$$\text{DCF} = \left(\frac{\text{number of misses (threshold} = t)}{\text{number of target trials}} + \frac{\text{number of false alarms (threshold} = t)}{\text{number of non-target trials}}\right) \times 100 \tag{3}$$

where 'misses' denotes the incorrect rejection (false negative) and 'false alarm' is the incorrect acceptance (false positive). The cost function can be evaluated at the threshold defined by t. The challenge evaluates the system on the basis of the DCF score (calculated by the challenge website after the system is submitted to it). Put simply, the lower the DCF, the better the system is at the speaker detection task.

Late fusion approach: In our approach, we combine the scores from the cosine distance and LDA by a linear relationship as defined by

$$\text{score} = \alpha \times \text{score1} + \beta \times \text{score2} \tag{4}$$

where score1 represents the cosine distance scoring and score2 represents the LDA scoring. See Fig. 1.

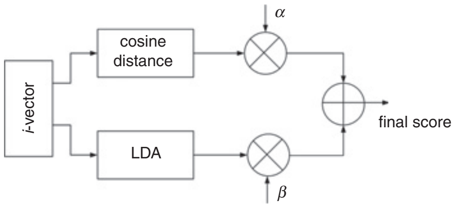

Fig. 1 *Unimodal late fusion*

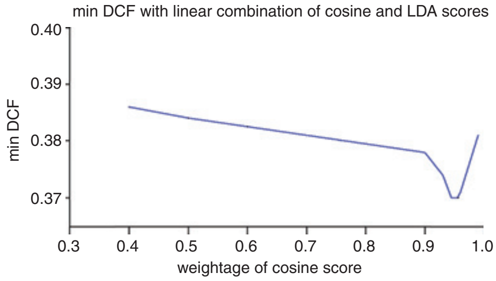

Fig. 2 *DCF against parameter 'α' (β = 1 – α)*

This approach is based on a concept of late fusion or fusion at the decision level. The counterpart of late fusion is early fusion which is fusion at the features level [5–7]. As presented in the literature, late and/or early fusion are widely popular for use in multimodal systems such as combining audio data with visual scores and/or features for speech recognition, multimodal biometric systems etc. [8, 9]. Generally speaking, the multimodal fusion can be categorised as feature level fusion, score level fusion and decision-level fusion. Feature level fusion is achieved by the concatenation of features from different modalities. Score level fusion is the linear combination of the scores for different modalities, e.g. taking the mean of the scores. The decision-level fusion is the logical decision fusion of the decision output for different modalities, e.g. by taking logical AND or logical OR of the decision from two systems. In this Letter, we are using the score level fusion for decision making on unimodal data. The motivation behind late fusion is the objective of taking advantage of the complementary information of the two techniques. To the best of our knowledge, this has not been reported previously for the *i*-vector paradigm. The *i*-vector challenge does not accommodate features other than *i*-vectors and thus early fusion is not applicable for this task.

Results: The *i*-vector challenge dataset is based on the previous data as used in NIST speaker recognition evaluations. Our approach presented in this Letter achieves a DCF of 0.370 better than the baseline 0.386 on the progress set. On the evaluation set (the challenge website showed the performance for the progress set of the test data, thus Fig. 2 is shown for the progress set of the test data; only the best score is shown for the evaluation set of the test data), our approach achieves a DCF of 0.363, again outperforming the baseline which is 0.378 (the best value for the DCF reported for the *i*-vector challenge at the time of writing this Letter is 0.240.). Empirical results show that the best results are obtained for $\alpha = 0.95$ as shown in Fig 2. β is equal to $1 - \alpha$.

Conclusion: In this Letter, we have presented a simple late fusion approach for the *i*-vector speaker recognition challenge. We have shown that the linear combination of scores from the cosine distance and the LDA results in outperforming the state-of-the-art baseline. The results are convincing and will motivate researchers to explore similar approaches for further improvements in the DCF. We also tested support vector machines (SVMs) with the Gaussian RBF kernel and submitted the score both in the stand-alone mode and fused with the baseline; however, the best DCF value achieved was 0.379, failing to outperform the LDA on this particular dataset. This leaves out LDA to be an optimal choice given that the LDA is much simpler than SVMs. Although these initial findings are very useful, there can be further improvements by combining the *i*-vector domain with an unsupervised features learning.

Acknowledgments: Hazrat Ali is supported by funding from the Erasmus Mundus STRoNGTiES grant. We also acknowledge Mathworks, Inc., for baseline scoring.

22 April 2014
doi: 10.1049/el.2014.1207
One or more of the Figures in this Letter are available in colour online.

Hazrat Ali, Artur S. d'Avila Garcez and Son N. Tran (*Department of Computer Science, City University London, Northampton Square, London, EC1V 0HB, United Kingdom*)

E-mail: engr.hazratali@yahoo.com

Xianwei Zhou and Khalid Iqbal (*School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, People's Republic of China*)

Hazrat Ali: Also with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China

References

1 'The 2013–2014 Speaker Recognition i-vector Machine Learning Challenge', 2014. [Online]. Available at https://www.ivectorchallenge.nist.gov, accessed 1 March 2014

2 Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., and Ouellet, P.: 'Front-end factor analysis for speaker verification', *IEEE Trans. Audio Speech Lang. Process.*, 2011, **19**, (4), pp. 788–798

3 Dehak, N., Kenny, P., Glembek, O., Dumouchel, P., and Burget, L.: 'Support vector machines and joint factor analysis for speaker verification', *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Taipei, Taiwan, April 2009, pp. 4237–4230

4 Balakrishnama, S., and Ganapathiraju, A.: 'Linear discriminant analysis; a brief tutorial'. [Online]. Available at http://www.music.mcgill.ca/~ich/classes/mumt611_07/classifiers/lda_theory.pdf

5 Gunes, H., and Piccardi, M.: 'Affect recognition from face and body: early fusion vs. late fusion'. IEEE Int. Conf. Systems, Man and Cybernetics: 2005, Vol. 4, pp. 3437–3443

6 Snoek, C.G.M., Worring, M., and Smeulders, A.W.M.: 'Early versus late fusion in semantic video analysis'. 13th Annual ACM Int. Conf. Multimedia, Waikoloa, HI, USA, October 2005, pp. 399–402

7 Kittler, J., Hatef, M., Duin, R.P., and Matas, J.: 'On combining classifiers', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1998, **20**, (3), pp. 226–239

8 Lip, C.C., and Ramli, D.A.: 'Comparative Study on Feature, Score and Decision Level Fusion Schemes for Robust Multibiometric Systems,' in 'Frontiers in computer education, advances in intelligent and soft computing', (Springer, 2012), pp. 941–948

9 Dass, S.C., Nandakumar, K., and Jain, A.K.: 'A principled approach to score level fusion in multimodal biometric systems', Audio Video-Based Biometric Person Authentication, *Lect. Notes Comput. Sci.*, 2005, Vol. 3546, pp. 1049–1058