Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 126 (2018) 1534–1540

Procedia Computer Science

www.elsevier.com/locate/procedia

22nd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

I-vector Extraction for Speaker Recognition Based on Dimensionality Reduction

Noor Salwani Ibrahim[a], Dzati Athiar Ramli[a]*

[a]*School of Electrical and Electronic, Universiti Sains Malaysia Engineering Campus, Nibong Tebal, Pulau Pinang 14300, Malaysia*

Abstract

In the domain of speaker recognition, many methods have been proposed over time. The technology for automatic speaker recognition has now reached a good level of performance but there is still need of improvement. In this paper, a new low-dimensional speaker- and channel-dependent space is defined using a simple factor analysis also known as i-vector. This space is named the total variability space because it models both speaker and channel variabilities. The i-vector subspace modelling is one of the recent methods that have become the state of the art technique in this domain. This method largely provides the benefit of modelling both the intra-domain and inter-domain variabilities into the same low dimensional space. In this study, 2656 syllables bio-acoustic signals from 55 species of frog taken from Intelligent Biometric Group, USM database are used for frog identification system. Parameters of the system are initially tuned such as Universal Background Model (UBM) size (32, 64 and 128 Gaussians) and i-vector dimensionality (100, 200 and 400 dimensions). To the end, we assess the effect of the parameter tuned and record the computation time. We observed that, the accuracy for smaller UBM size and higher i-vector dimensionality outperforms others with result of 91.11% is achieved. From this research, it can be concluded that UBM size and i-vector dimensionality effect the accuracy of frog identification based on i-vector.

Keywords: Bob Spear toolbox, I-vectors, Dimensionality Reduction, UBM size, Frog Identification.

* Corresponding author. Tel.: +6-004-599-5999 ; fax: +6-004-594-1023.
E-mail address: dzati@usm.my

10.1016/j.procs.2018.08.126

1. Introduction

Speaker recognition is the identification of a person or species (for animal) from characteristics of voices. It is also called as voice recognition. Automatic speaker recognition is the use of a machine to recognize a person's identity from the characteristics of his voice. The technology employed in this field has now reached a good level of performance due to the success of new paradigms such as session variability modelling [1, 2] and total variability (i-vectors) modelling [3]. Furthermore, it benefited from improvements in channel compensation [4, 5, 6] and noise reduction [7, 8] techniques. As a matter of fact, the NIST speaker recognition evaluation (SRE) series [9] have seen a record in their number of participants in the last editions. The same trend was observed at the speaker recognition evaluation in mobile environments.

Speech or sound has always been the ideal method of communication for humans or between animals, but for making it efficient while interacting with machines has always been a challenge. Speech processing has really evolved in the last few decades owing to the advances in the methods of feature extraction and dimensionality reduction. The main hurdle in recognizing sound is that each speaker has his or her unique way of speaking, accent, pitch, rhythm, emotional state, etc. and there are differences even in the physical characteristics like vocal tract shapes or other sound production organs. These differences pose difficulty in extracting the similar traits required for a particular recognition application like language, accent or speaker from various speech utterances.

In spite of all these difficulties it is important to develop speaker recognition applications because of their usability in various domains. Nowadays more and more sound based services are facing a paradigm shift of moving towards automated systems. These systems are more capable of handling the loads during peak times when not many manual options are available. But due to low accuracy these systems are not used in critical environments. Various methods for speaker processing have been proposed over time but only a few of them have really been put to practical use.

Among the existing systems, total variability model or i-vector approach originally used for speaker recognition [3], has been experimented for frog identification system in this research. The i-vector system has many data-driven components for which training data needs to be selected [10]. It would be tempting to train some of the hyper-parameters on a completely different out-of-set-data and leave only the final parts (training and testing a certain sound) to the trainable parts. The lack of resources can cause the performance not very good compared to the traditional Gaussian Mixture Models (GMM) approach [11,12]

Inspired by the earlier use of Joint Factor Analysis (JFA) speaker factors directly as features for Support Vector Machine (SVM) classification, Dehak et al. [3] 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. [3] that the channel space of JFA contains information that can be used to distinguish between speakers. The total variability factors are an 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 [3,13].

1.1. I-vector

A simpler model for speaker recognition has been introduced in [3], which gets rid of the distinction between speaker and channel variability subspaces, and models both in a common constrained low dimensional space, referred to as the "total variability space". Speaker or session variability is the variability exhibited by a given speaker from one recording session to another. This type of variability is usually attributed to channel effects although this is not strictly accurate since intra-speaker variation and phonetic variation are also involved [14]. In this approach, a speech segment is represented by a low-dimensional "identity vector" (i-vector for short) extracted by Factor Analysis. The i-vector approach has become state-of-the-art in the speaker verification field [3] and, in this work, we show that it can be successfully applied also to frog identification. The approach provides an elegant

way of reducing high-dimensional sequential input data to a low-dimensional fixed-length feature vector while retaining most of the relevant information [15]. The main idea is that the session- and channel-dependent supervectors of concatenated Gaussian Mixture Model (GMM) means can be modelled as

$$s = m + Tw \quad (1)$$

where m is the session- and channel-independent component of the mean supervector, T is a matrix of bases spanning the subspace covering the important variability (both speaker- and session-specific) in the supervector space, and w is a standard normally distributed latent variable. For each observation sequence representing an utterance, our i-vector is the Maximum A Posteriori (MAP) point estimate of the latent variable w. Our i-vector extractor training procedure is based on the efficient implementation suggested in [16].

The contribution of this study is to evaluate the result from the factors affecting the i-vector based on frog sound identification. We study this from parameter perspectives where we evaluate and analyse how the various i-vector extractor parameters, such as the Universal Background Model (UBM) size and i-vector dimensionality, affect frog detection accuracy. UBM size refers to Gaussians component that is corresponding adapted component in the speaker model. I-vector dimensionality is equal to the "rank" of eigenmatrix. Based on Huang et al. [17], a larger i-vector dimension would not give much improvement to the classification performance but it greatly increased the computation effort. In Xu et al. [10] discussed by reducing the computation will allow effective use of i-vector in more applications. In this research, the computation time recorded is to investigate either both factors affect the computation or not and the next direction for next research is determined.

The remainder of the paper is as follows. Section 2 describes the experimental system methodology including the total variability modelling approach and scoring. Section 3 presents the results obtained for the system. Section 4 includes conclusions and future work.

2. Methodology

Figure 1 shows the block diagram of the method used in this work. The i-vector system consists of two main parts: front-end and back-end. The former consists of cepstral feature extraction and UBM training, whereas the latter includes sufficient statistics computation, training of the T-matrix, i-vector extraction, dimensionality reduction and scoring.

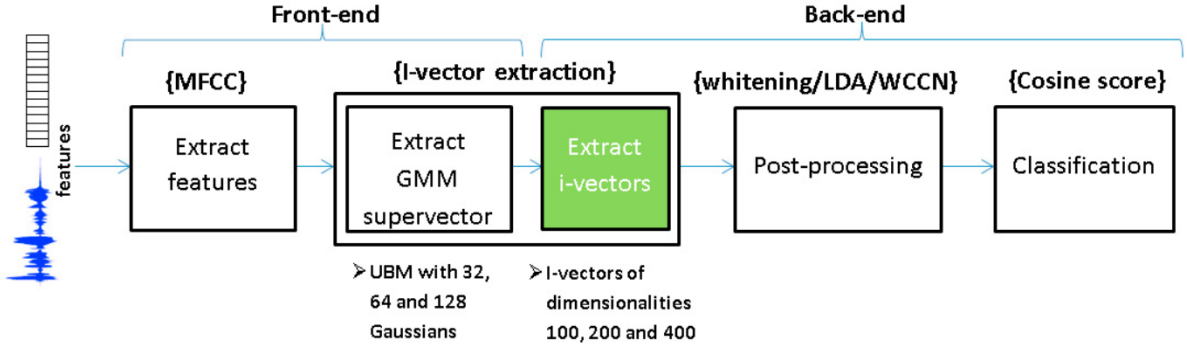


Fig. 1. Block diagram for frog identification system experiment

2.1 Feature Extraction

First, simple energy-based voice activity detection (VAD) is performed to discard the non-frog-sound parts. Energy based VAD technique is used where frame-level energy values are first computed followed by data normalization and finally, the detection of voice activity process. The class with the higher mean is considered as

frog sound, and the corresponding sound segments are hence kept before being smoothed. Second, MFCC and log energy features together with their first- and second order derivatives are computed over 20 ms hamming windowed frames every 10 ms. Frog sound activity detection is then applied and the frog sound is normalized following a standard normal distribution.

2.2 I-vector extraction

For each utterance, the corresponding feature sequence is finally converted to an i-vector using an i-vector extractor based on a GMM with three different UBM size components trained on pooled features from all samples included in our training data. The three UBM size as stated in Table 1 which are 32, 64 and 128 components.

2.3 Post-processing

As the i-vectors modelling contains speaker and channel variability information simultaneously in one space, carrying out channel compensation technique in the total factor space is required in order to remove the nuisance effects. The channel compensation approaches play a major role in the i-vector speaker recognition systems. Channel compensation is therefore necessary to make sure that test data obtained from different channels can be properly scored against the speaker models. For channel compensation to be possible, the channel variability has to be modelled explicitly. Before calculating the verification scores, whitening, Linear Discriminant Analysis (LDA) and Within-Class Covariance Normalization (WCCN) projections were performed for channel compensation. We used the same data set for training the total variability matrix to estimate the LDA and WCCN matrices.

As the extracted i-vectors contain both intra- and between-accent variations, the aim of dimensionality reduction is to project the i-vectors onto a space where between-accent variability is maximized and intra-accent variability is minimized. Three different dimensions that have been experimented in this research are 100, 200 and 400 dimensions. In summary, we optimized the i-vector parameters as Table 1 to experiment and evaluate the result.

Table 1. The i-vector system's parameter experiment

i-vector parameter	Range and optima
UBM size	32, 64, 128
i-vector dimensionality	100, 200, 400

2.4 Scoring

Finally, the identification result from the system is given by calculating the similarity score is computed. The simplest and fast scoring function i.e. the cosine distance is calculated between the i-vectors from a speaker model and the i-vector from the test segment. The decision and evaluation process is then computed and the system performance is then presented by accuracy (%), CMC curves and detection error trade-off (DET).

3. Result and discussion

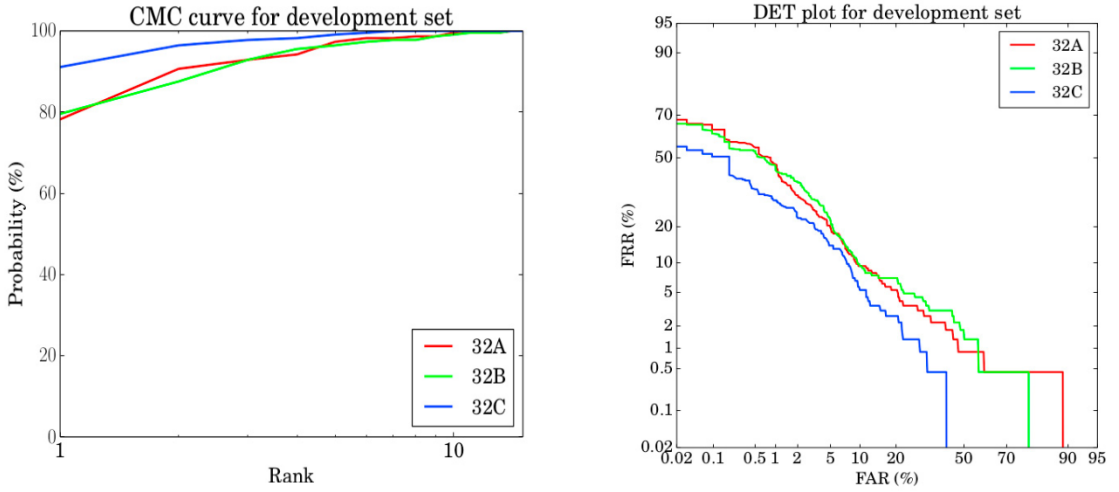
The experiments have been done using Spear toolbox in Ubuntu operating system. Spear is an open source and extensible toolbox for state of the art speaker recognition. This toolbox is built on top of Bob, a free signal processing and machine learning library. 2656 syllables bio-acoustic signals from 55 frog species taken from Intelligent Biometric Group, Universiti Sains Malaysia database have been used for evaluation.

Table 2 shows the performance of the i-vector system for selected i-vector dimension, selected UBM size and computation time. As presented in Table 2, the best accuracy is given by i-vector dimensionality with 400 dimensions with 91.11%, 84.89% and 79.11% respectively for the three different UBM size. We found that accuracy improves with the increase of i-vector dimensionality. Furthermore, our results showed that the UBM with smaller size outperforms larger UBM. Apart from that, the computation time result shows that it takes longer time to process when the i-vector dimensionality in increase and UBM size is bigger.

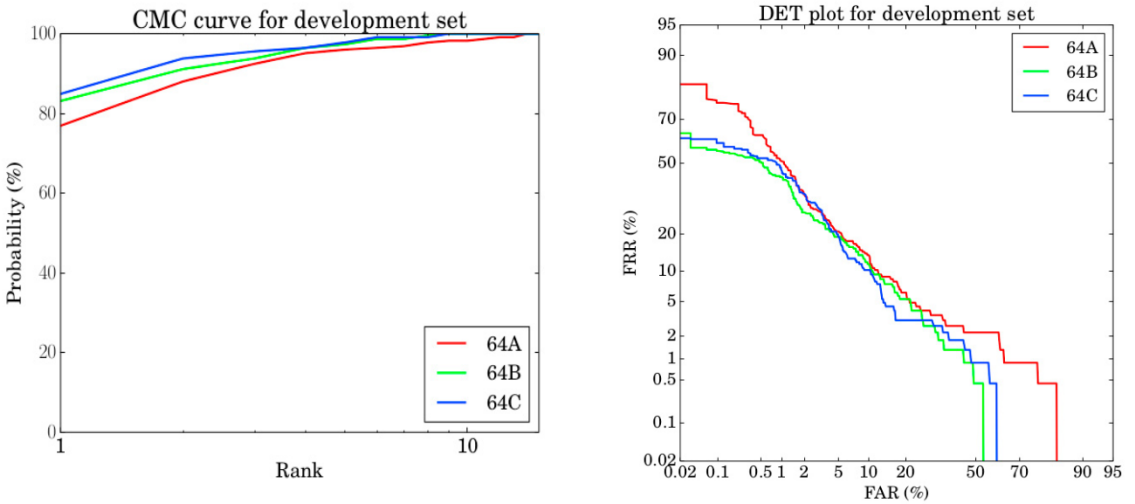
Table 2. Performance of the i-vector system for selected i-vector dimensions (accuracy in % and processing time in second)

i-vector dimensionality	Performance (Accuracy % and computation time)					
	32 Gaussian	Time(s)	64 Gaussian	Time(s)	128 Gaussian	Time (s)
A. 100	78.22	15091	76.89	19322	72.44	22997
B. 200	79.56	21249	83.11	22325	73.33	28611
C. 400	91.11	23523	84.89	26094	79.11	34981

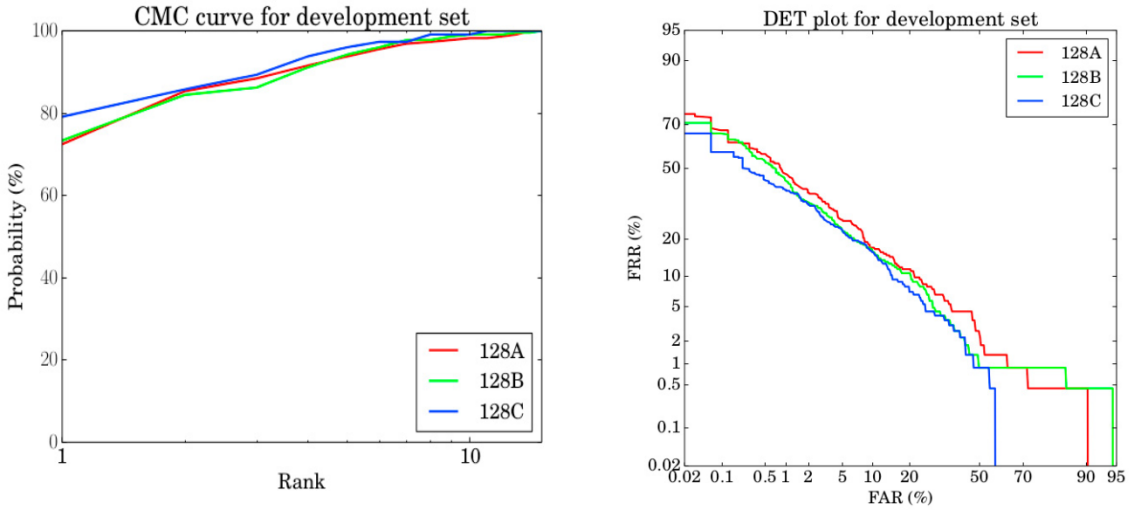
Fig. 2 shows the CMC curves and DET plots of; a) 32 Gaussian UBM size, b) 64 Gaussian UBM size and c) 128 Gaussian UBM size. Those plots confirm the same conclusion as given above as accuracy improves with increase i-vector dimensionality and UBM with smaller size gives better performance compared to larger UBM size.



a. 32 Gaussian UBM size



b. 64 Gaussian UBM size



c. 128 Gaussian UBM size

Fig. 2. CMC curves and DET plots for 3 different UBM size

4. Conclusion

In this paper, we studied how the i-vector extractor parameter like UBM size and i-vector dimensionality affects frog identification detection accuracy. Regarding parameters, highest accuracy was achieved using UBMs with 128 Gaussians and i-vector dimensionality of 400. These are similar to those reported in general speaker recognition literature. Regarding the data, we found that the choice of the UBM training data is most critical part, followed by i-vector dimensionality. This is understandable since the earlier system components affect the quality of the remaining steps.

For further research, we suggest to investigate the effect for higher i-vector dimensionality and higher UBM size. For this research, we are not doing that due to long computation time because we are using small size of frog database. Next studies we can reduce computation time by investigating other factor affecting and add more data for further investigate this experiment effect.

Acknowledgements

This work was sponsored and supported by Research University Grant, Universiti Sains Malaysia (1001.PELECT.8014057).

References

[1] Patrick Kenny, G. Boulianne, Pierre Ouellet and Pierre Dumouchel. (2007) "Joint factor analysis versus eigenchannels in speaker recognition." *IEEE Trans. on Audio, Speech, and Language Processing* **15(4)**:1435–1447.

[2] Robert J. Vogt and Sridha Sridharan. (2008) "Explicit modelling of session variability for speaker verification." *Computer Speech & Language* **22(1)**:17–38.

[3] Najim Dehak, Patrick Kenny, Reda Dehak, Pierre Dumouchel and Pierre Ouellet. (2011a) "Front-end factor analysis for speaker verification." *IEEE Trans. on Audio, Speech, and Language Processing* **19**:788–798.

[4] Alex Solomonoff, W. M. Campbell, Ian Boardma. (2005) "Advances in channel compensation for svm speaker recognition." *in IEEE ICASSP* **1**: 629–632.

[5] Daniel Garcia-Romero and Carol Y. Espy-Wilson. (2011) "Analysis of i-vector length normalization in speaker recognition systems." *in INTERSPEECH*: 249–252.

[6] Najim Dehak, Reda Dehak, Patrick Kenny, Niko Brummer, Pierre Ouellet, Pierre Dumouchel. (2009) "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification." *in INTERSPEECH*: 1559–1562.

[7] Ji Ming, Timothy J. Hazen, James R. Glass and Douglas A. Reynolds. (2007) "Robust speaker recognition in noisy conditions." *IEEE Trans. on Audio, Speech, and Language Processing* **15(5)**: 1711–1723.

[8] Yun Lei, Lukas Burget, and Nicolas Scheffer. (2013) "A noise robust ivector extractor using vector taylor series for speaker recognition." *in IEEE ICASSP*: 6788–6791.

[9] Craig S Greenberg, Vincent M Stanford, Alvin F Martin, Meghana Yadagiri, George R Doddington, John J Godfrey, Jaime Hernandez. (2013) "The 2012 NIST Speaker Recognition Evaluation." *in INTERSPEECH:* 1971-1975.

[10] Longting Xu, Kong Aik Lee, Haizhou Li, Zhen Yang. (2018) "Generalizing I-Vector Estimation for Rapid Speaker Recognition." *IEEE/ACM Transactions on Audio Speech and Language Processing* **26(4)**:749-759.

[11] Hamid Behravan, Ville Hautamaki, Tomi Kinnunen. (2015) "Factors affecting i-vector based foreign accent recognition: A case study in spoken Finnish." *Speech Communication* **66**:118–129.

[12] Giorgio Biagetti, Paolo Crippa, Laura Falaschetti, Simone Orcioni and Claudio Turchetti. (2017) "An investigation on the accuracy of truncated DKLT representation for speaker identification with short sequences of speech frames." *IEEE Transactions on Cybernetics* **47 (12)**:4235-4249.

[13] Omid Ghahabi and Javier Hernando. (2017) "Deep Larning Backend for Single and Multisession i-Vector Speaker Recognition." *IEEE/ACM Transactions on Audio Speech and Language Processing* **25 (4)**:807-817.

[14] Giorgio Biagetti, Paolo Crippa, Laura Falaschetti, Simone Orcioni and Claudio Turchetti. (2018) "Speaker identification in noisy conditions using short sequences of speech frames." *Smart Innovation, Systems and Technologies,* Springer, Cham **73**: 43-52.

[15] Patrick Kenny, Pierre Ouellet, Najim Dehak, Vishwa Gupta and Pierre Dumouchel. (2008) "A study of interspeaker variability in speaker verification." *IEEE Trans. Audio, Speech Lang. Process* **16(5)**: 980–988.

[16] Ondrej Glembek, Lukas Burget, Pavel Matějka, Martin Karafiát, and Patrick Kenny. (2011) "Simplification and optimization of i-vector extraction." *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. – Proc.:* 4516–4519.

[17] Zhen Huang, You-Chi Cheng, Kehuang Li, Ville Hautamaki, Chin-Hui Lee. (2013) "A blind segmentation approach to acoustic event detection based on I-vector." *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH* August: 2282–2286.