In International Journal of Speech Technology, volume 18(4), pages 529–546, 2015

i-Vectors in Speech Processing Applications: A Survey

Pulkit Verma · Pradip K. Das

Abstract In the domain of speech recognition many methods have been proposed over time, like Gaussian mixture models (GMM), GMM with universal background model (GMM-UBM framework), joint factor analysis, etc. i-Vector subspace modeling is one of the recent methods that has become the state of the art technique in this domain. This method largely provides the benefit of modeling both the intra-domain and inter-domain variabilities into the same low dimensional space. In this survey, we present a comprehensive collection of research work related to i-vectors since its inception. Some recent trends of using i-vectors in combination with other approaches are also discussed. The application of i-vectors in various fields of speech recognition, viz speaker, language, accent recognition, etc. is also presented. This paper should serve as a good starting point for anyone interested in working with i-vectors for speech processing in general. We then conclude the paper with a brief discussion on the future of i-vectors.

Keywords Speech processing, Feature extraction, JFA, Factor analysis, i-Vectors, PLDA

This is a post-peer-review, pre-copyedit version of an article published in International Journal of Speech Technology. The final authenticated version is available online at: https://doi.org/10.1007/s10772-015-9295-3
Two links updated in March 2021 (on page 13 and 14), as the links used in the original paper became invalid/dead.

Pulkit Verma[1]
v.pulkit@iitg.ernet.in

Pradip K. Das[1]
pkdas@iitg.ernet.in

[1]Department of Computer Science & Engineering,
Indian Institute of Technology Guwahati,
Guwahati, Assam 781039, India

1 Introduction

Speech has always been the ideal method of communication for humans, 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 speech is that each speaker has his or her unique way of speaking, accent, pronunciation, 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 difficulties 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 speech recognition applications because of their usability in various domains. Nowadays, more and more voice based services are facing a paradigm shift of moving towards automated systems. These systems are more capable of handling 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 speech processing have been proposed over time but only a few of them have really been put to practical use. Some of them included the use of Gaussian Mixture Models (GMMs) for speaker verification (Reynolds, 1992), Universal Background Model (UBM) along with GMMs(Reynolds et al., 2000), and more recently Joint Factor Analysis (JFA) (Kenny et al., 2007a).

All these methods follow the same pattern but differ in the implementation. The general framework of any speech recognition application can be explained using a speaker verification system as shown in Fig. 1. Applying

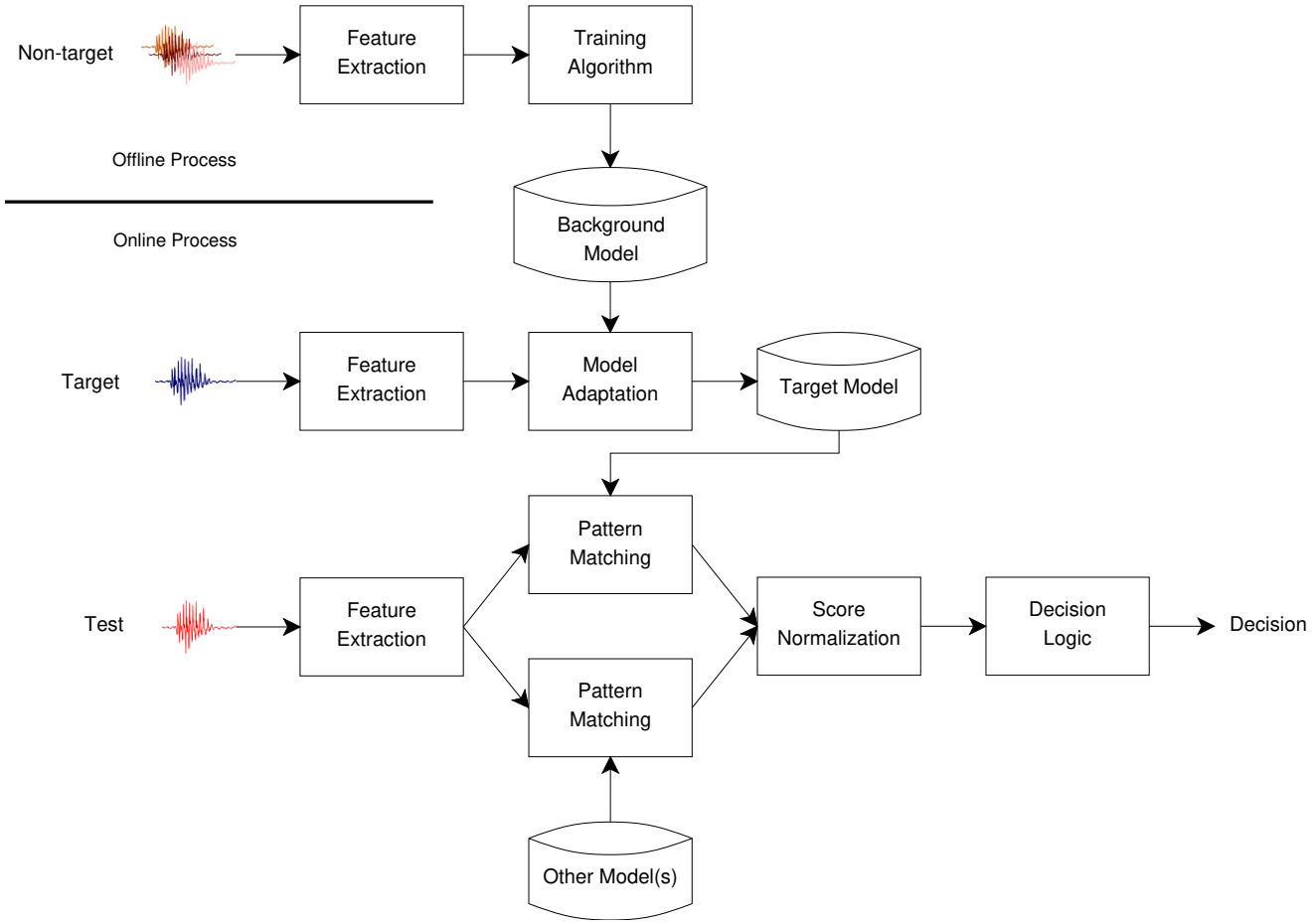

Fig. 1 Flowchart for a generic Speaker Recognition Process

it to any other domain requires changing the suitable parameters while giving speech data as input to the system. This generic approach can be subdivided into two main steps, speaker enrollment and speaker verification.

During the speaker enrollment process, a background model is generated using the data collected from non-target utterances. In this paper, for nearly all the approaches this model will be UBM. Using this background model, a target model is generated using the target utterances by adapting the background model according to the target data. Now this target model will act as the single point of reference for the pattern recognition algorithms. Whenever any test utterance is given as input to the system, features are extracted from it and pattern matching algorithms are applied on it using one or more kinds of target models. The resultant similarity score is then normalized and final decision is taken after applying some decision logic to this normalized score.

An important aspect of using any of the techniques mentioned in this paper is the feature extraction step. The features used for this purpose must fulfill some criteria so as to give better performance. According to Wolf (1972), some of the important characteristics of these feature vectors are that they should:

- be less susceptible to environmental noise
- occur frequently and naturally in normal speech
- not be affected by health of the speaker
- be easily measurable
- be difficult to copy by impostors

Due to these reasons, generally Mel-frequency cepstral coefficients (MFCCs) (Picone, 1993) are used as feature vectors for speech recognition applications. A simple flowchart depicting the MFCC extraction process is depicted in Fig. 2.

The concept of i-vectors which was initially proposed for speaker verification in Dehak et al. (2011b) has become quite popular in other speech processing applications. This approach tries to improve the JFA by combining the inter and intra domain variability and modeling it in same low dimensional total variability space.

This paper is a survey on i-vectors in the domain of speech recognition applications. Since i-vectors are developed by making modifications in JFA, it is essential

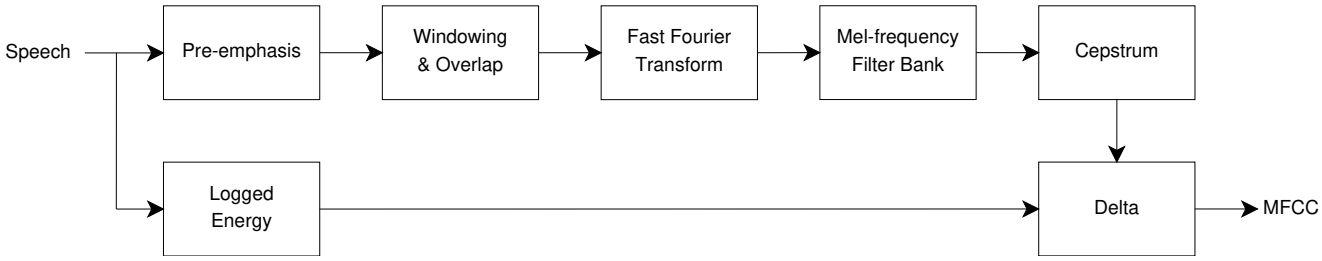

Fig. 2 Flowchart for MFCC extraction (Dehak and Shum, 2011)

to understand JFA properly before moving on to the explanation of i-vectors. While i-vectors were originally proposed for speaker verification, they have been used extensively in text-dependent speaker verification applications (Larcher et al., 2012), language recognition (Martínez et al., 2011), speaker diarization (Silovsky and Prazak, 2012), etc.

The rest of the paper is organized as follows: After the Introduction, Sect. 2 focuses on JFA as it is important for understanding i-vectors; Sect 3explains the i-vectors in detail and some of its applications in various domains; Sect. 4 discusses some other approaches which are used along with i-vectors to improve the speech recognition; Sect. 5 describes some of the toolkits providing i-vector support; Sect. 6 discusses the future of i-vector technique; Sect. 7 concludes the work.

2 Joint Factor Analysis

In Joint Factor Analysis (JFA) (Kenny et al., 2007a,b, 2008), a supervector M is used to represent any speech utterance. M contains speaker and channel dependent supervectors.

Let C be the number of components in UBM and F be the dimension of acoustic feature vectors.

Now for a given utterance, we concatenate the F-dimensional GMM mean vectors to get the supervectors of dimension CF. For a particular speaker, we assume that the speaker and channel dependent supervector M is generated by the vector sum of speaker-dependent supervector and channel dependent supervector. Also, these speaker and channel supervectors are distributed normally and are statistically independent.

$$M = s + c \tag{1}$$

where, M is speaker and channel dependent supervector, s is speaker dependent supervector, and c is channel dependent supervector.

Both the speaker and channel factors are decomposed into low dimensional set of factors. Each of these low dimensional factors operate along the principal dimensions of the corresponding component.

We assume that the distribution of speaker dependent supervector s has a hidden variable description of the form:

$$s = m + Vy + Dz \tag{2}$$

where, s is speaker dependent supervector, m is $CF \times 1$ speaker and channel independent supervector (from UBM), V is eigenvoice matrix (rectangular matrix of low rank), D is $CF \times CF$ residual diagonal matrix, y is a vector representing speaker factors, z is a normally distributed CF-dimensional random vector representing speaker specific residual factors.

The decomposition of a factor into lower dimensional factors can be illustrated by the following decomposition of speaker dependent component of speaker factor:

$$V * y = \begin{bmatrix} | & | & & | \\ v_1 & v_2 & ... & v_N \\ | & | & & | \end{bmatrix} * \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_N \end{bmatrix} \tag{3}$$

where, V is eigenvoice matrix, y is lower dimensional vector representing speaker (or eigenvoice) factors. Each speaker factor y_i controls the corresponding v_i.

We assume that the distribution of channel dependent supervector c has a hidden variable description of the form:

$$c = Ux \tag{4}$$

where, c is channel dependent supervector, U is eigenchannel matrix (matrix of low rank), x is a normally distributed random vector representing channel factors.

With each mixture component i, a diagonal covariance matrix Σ_i is associated. The variability in acoustic observation vectors is not modeled by either (1) or (4).

We now define a $CF \times CF$ super-covariance matrix Σ whose diagonal is obtained by concatenating all the covariance matrix Σ_i. The variability that is not modeled by s and c is modeled by Σ.

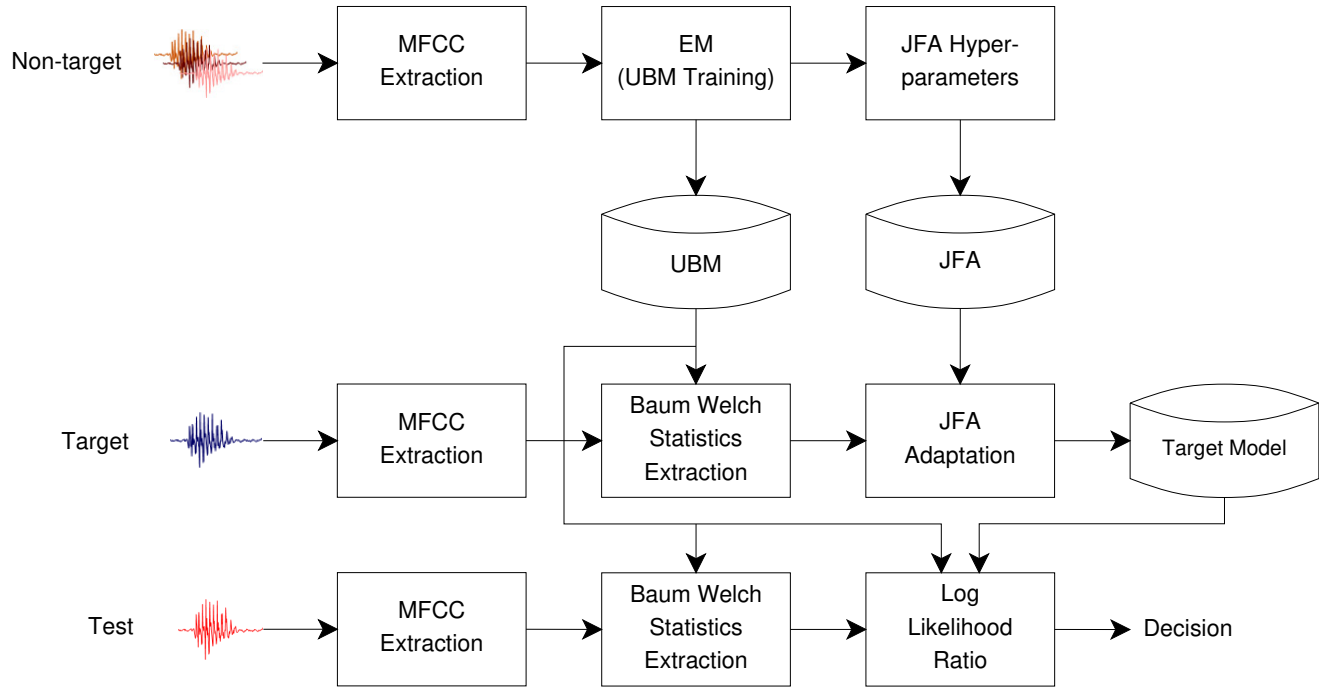

Fig. 3 Flowchart for Joint Factor Analysis (Dehak and Shum, 2011)

2.1 JFA Working

Once the MFCCs for all utterances in the training set are extracted, estimation maximization is applied to generate the UBM. JFA hyper-parameters are also extracted as explained in Kenny (2005). For the target dataset, once the MFCCs are extracted, the zeroth and the first order Baum–Welch statistics are extracted. Now using the JFA hyper-parameters extracted earlier adaptation is done on the data to generate the target model.

For performing the recognition of any test utterance, MFCCs are extracted followed by computation of the Baum–Welch statistics. Now using the target model and these statistics values the log likelihood is calculated for each speaker. The maximum value is chosen as the recognized speaker. Fig. 3 explains this process in detail.

It should be noted that log-likelihood calculation is not the only scoring method used with JFA and its choice may vary across applications. Various such scoring methods used with JFA for speaker recognition are compared in Glembek et al. (2009). It was deduced that though the performance of the various techniques does not vary significantly, the real difference was in the speed of scoring. Linear scoring is found to be fastest of all the scoring techniques as the channel compensation is calculated only once per speech utterance.

The various algorithms and detailed description of JFA can be found in Kenny (2005).

2.2 JFA in Speech Applications

Yin et al. (2006) discuss various ways of performing speaker adaptation when applying factor analysis for speaker recognition. They showed that instead of using one enrollment utterance per speaker, it is better to use 8 utterances per speaker.

JFA based techniques have also been successfully used in language recognition evaluation (LRE) task (Jancik et al., 2010; Brümmer et al., 2009). The basic idea of JFA when applied to language recognition, takes into account the inter-class variability and and inter-session variability. Inter-class variability models the differences in languages whereas the inter-session variability (or channel variability) models the differences in utterances of same languages caused due to channel and/or speaker. While testing an utterance, the models of all languages are adapted to the channel of that utterance using maximum a posteriori (MAP) probability or maximum likelihood (ML) parameter estimation.

3 i-vector Approach

In JFA, it was assumed that the channel factors will only model the channel effects, but Dehak (2009) observed that the channel dependent supervector also models the speaker features. To take this finding into account, a new approach was proposed where there was no distinction between channel and speaker variabilities.

A new low dimensional total variability space T was introduced to account for both the variabilities, where

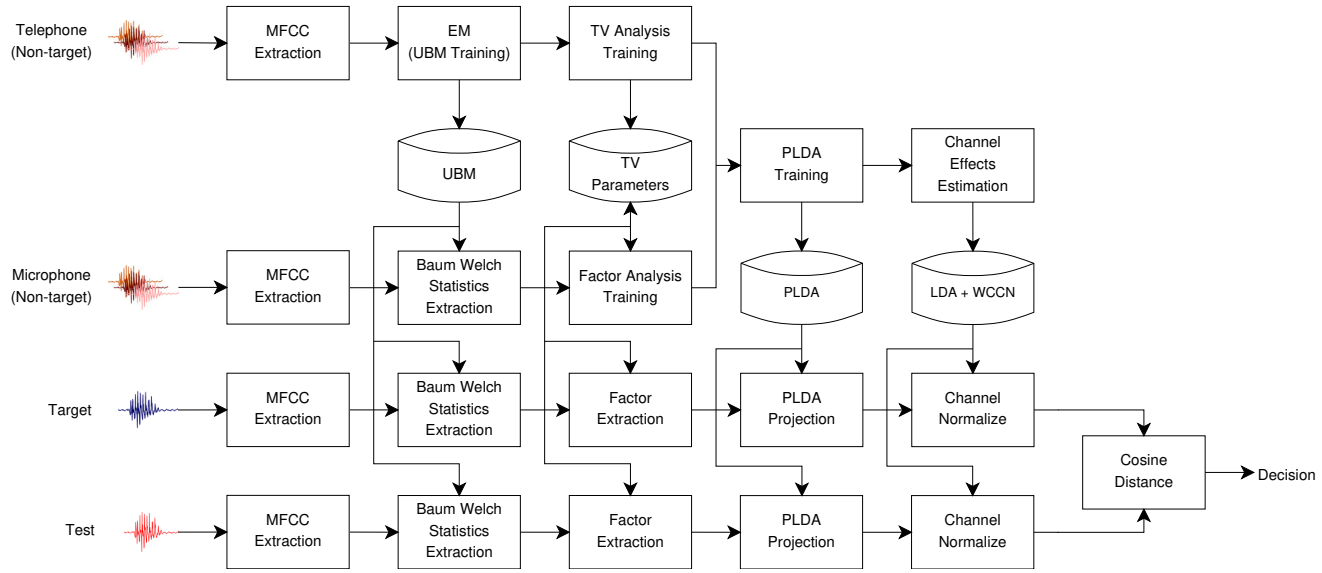

Fig. 4 Flowchart for channel blind i-vector based approach (Dehak and Shum, 2011)

M is given by:

$$M = m + Tx \tag{5}$$

where, m is the UBM supervector (speaker and channel independent supervector), x is a normally distributed random vector in this space and T is a low ranked rectangular matrix. The factors of x are also called as total factors. The new vectors are known as identity vectors or i-vectors.

In this approach, it is assumed that M is distributed normally with TT' as its covariance matrix and m as its mean vector. Here, the total variability matrix T is obtained in the same manner as that of V, but it is assumed that the utterances of the same speaker are produced by different speakers unlike in the process of training eigenvoice matrix V.

In order to bring down computation, the channel compensation is done in total factor space as the dimensionality of i-vectors is lower as compared to GMM supervectors. Hence, the computation cost is much less as compared to JFA.

One of the additional advantage of this approach is that supervised training is not needed in this model unlike JFA and GMM-UBM.

Many applications of i-vectors are possible, some of which are:

1. Language Recognition
2. Accent/Dialect Recognition
3. Speech Diarization
4. Acoustic Event Detection

3.1 i-vector in Speaker Recognition

Since i-vector technique was introduced for the speaker verification task, most of its applications are aimed at solving this problem. In the literature, it can be seen that many researchers have modified either the feature set or the normalization technique (both length and score) to improve the performance of i-vector based systems. This section emphasizes on some of the proposed methods which uses i-vectors in the speaker verification scenario.

Most of the speaker recognition systems are trained on either telephone speech or microphone speech. Dehak et al. (2011a) proposed a channel-blind system which is conditioned on both telephone and microphone speech. To project both types of data on same space, PLDA was used. And for the hyper-parameter training, first μ, V, and Σ were trained on telephone data assuming $U = 0$ (EM algorithm was used). Then U was trained on microphone data assuming μ, V, and Σ are fixed (Maximum a posteriori probability (MAP) estimate was used). This approach is explained in detail in Fig. 4. This is the general template for nearly all the approaches that involve i-vectors. In the literature, there are reported works (Glembek et al., 2011; Aronowitz and Barkan, 2012) where efficient computations of i-vectors are presented. There are several other variations of this general approach which are listed below:

1. Using different features instead of MFCCs.
2. Using different feature or score normalization technique instead of PLDA, LDA, and WCCN.
3. Modifying the statistics calculation.
4. Using different scoring technique instead of Cosine Distance scoring.

5. Changing some inherent assumption of Front End Factor Analysis.

All these methods try to alter the standard approach so as to improve certain aspects of the recognition system.

In Senoussaoui et al. (2010), a modified version of cross channel condition proposed in Kenny et al. (2008) was used, where the two gender-dependent matrices T and T' were estimated. T was estimated using only telephone data, while T' was estimated using only microphone data. The supervectors M associated with these telephone and microphone recordings were then combined to produce new feature extractor that can be used for both microphone and telephone speech. The way in which telephone speech could be added in i-vector space to compensate the channel effects with linear discriminant analysis (LDA) and within class covariance normalization (WCCN) was also demonstrated.

The UBMs with full covariance matrices can be used effectively in i-vector systems. In the normalization of first order Baum–Welch statistics it plays an important role. The work by Matejka et al. (2011) also showed that Gaussian-PLDA (G-PLDA) is as effective as heavy tailed-PDA (HT-PLDA) for scoring. Similar results were reported by Garcia-Romero and Espy-Wilson (2011). It should be noted that HT-PLDA is similar to G-PLDA with a minor difference that in HT-PLDA the priors in generative model are assumed to have Student's t distribution (heavy tailed priors) instead of Gaussian distribution assumed in G-PLDA. They used length normalization with G-PLDA to achieve performance close to HT-PLDA, which is much more complicated than G-PLDA. The reason for improved efficiency was attributed to the matrices of low rank involved in the computation of log-likelihood ratios.

In speaker recognition tasks, probabilistic linear discriminant analysis (PLDA) is commonly used to handle the speaker and session variability. One major reason favoring PLDA as explained by Jiang et al. (2012) is that dimensionality reduction of feature vector is done two times, once during the standard i-vector extraction and second time while applying the PLDA model. Hence, in supervector space we can afford to keep the full dimensionality of i-vectors thereby avoiding loss of information. Due to this reason, this work proposed the use of uncompressed i-vector (termed as i-supervector) when using PLDA for speaker recognition. A main characteristic of i-supervector is that it performs better without applying any feature or score normalization. A major drawback of using i-supervector was the problem faced in inverting large matrices which was efficiently overcome by dividing it into small blocks.

More domain specific modifications of i-vector-PLDA technique include multichannel simplified PLDA for the case when i-vectors are generated from different recording conditions (Villalba and Lleida, 2013), use of weighted-LDA (Kanagasundaram et al., 2012a), mixture of PLDA models for gender independent speaker recognition (Senoussaoui et al., 2011), supervized mixture of PLDA Models (Simonchik et al., 2012), PLDA based on restricted Boltzmann machines (RBM) (Novoselov et al., 2014), etc.

3.1.1 Mismatched Length

Sarkar et al. (2012) pointed out that having mismatched length of utterances for training and testing may pose problems for speaker verification tasks solved using i-vectors. From experiments it was shown that even when the test utterances are short, it would be better if the training is done on long utterances.

Kenny et al. (2013) demonstrated that i-vector with PLDA is also efficient in the case where utterances vailable for enrollment and testing tasks are of variable lengths. The utterances used in this work ranged from 3s to 60s. While the i-vectors extracted for short utterances are less reliable, the reliability increases significantly when longer utterances are used. The standard i-vector approach considers all point estimates of i-vectors as equally reliable hence cannot be used here. In the proposed approach, the uncertainty associated with i-vector data selection plays an important role in the performance of speaker extraction. This is then quantified and propagated to PLDA classifier. For achieving this, observation noise while extracting i-vectors is modeled using channel factors. Also the zero and first order Baum–Welch statistics are scaled by a factor of 1/5, 1/3, or 1. This scaling is done both during the training and at run time. Length normalization is also used with uncertainty propagation to improve the overall performance.

Hasan et al. (2013) performs a more detailed analysis of this duration mismatch in utterances by analyzing the effect of this mismatch on i-vector length and phoneme distributions. It was observed that with the decrease in the length of the utterance, (i) the number of unique phonemes detected decreases logarithmically, and (ii) the length of the i-vector decreases non-linearly. To solve the problem of recognition in this scenario, it was assumed that the variability in duration is an additive noise. Three approaches were then proposed to compensate for duration variability; (1) multi-duration PLDA training, (2) using quality measure function to compensate the domain score, and (3) generating the short duration i-vectors synthetically in PLDA training. All

these methods provided sufficient improvements over the baseline system trained on full duration utterances.

3.1.2 Short Duration Utterances

The length of the utterance plays a major role in accuracy of all kinds of speech recognition systems. Low performance of these systems for short utterances is largely attributed to the low phoneme count in them. Behavior of techniques focusing on the total variability paradigm when training and testing utterances are shortened was analyzed by Kanagasundaram et al. (2011). Out of the four approaches compared, JFA and G-PLDA showed marginally better performance than LDA + WCCN or short duration nuisance attribute projection (SD-NAP) + WCCN based i-vector systems. This work was extended in Kanagasundaram et al. (2012b) to infer that HT-PLDA actually performs better than G-PLDA when subject to short duration telephone utterances. It also shows that for short utterances the performance of HT-PLDA is improved when score normalization is used.

In some speaker verification systems where it is important to stop the impostors from entering the system, it might be feasible to use short pass phrases as verification texts. In such cases as shown by Larcher et al. (2012), if the target speakers know the phonetic content of the training data, the error rate in speaker verification drops to 15.38% even if the impostor is also aware of this phonetic content. And in the best case where only the target speakers know about it, then error rate drops to 8.02%. The application domain of this finding is limited to only the cases where we can have the flexibility of sharing the phonetic content of training data with target speakers apriori.

Hautamäki et al. (2013) proposed one of the few methods which aimed at manipulating the statistics calculation in the standard i-vector approach to improve the performance of the system. For the short duration utterances, minimax estimation was used to calculate the first order statistics for i-vector extraction. This approach outperformed the baseline method using Baum–Welch statistics estimation significantly.

For short utterances, normalization techniques like source and utterance-duration normalized LDA (SUN-LDA) have been proposed by Kanagasundaram et al. (2013). Two other variations of SUN-LDA, namely *SUN-LDA pooled* and *SUN-LDA concat*, have also been proposed where pooling and concatenation, respectively are used with SUN-LDA as the normalization method. These methods are used to compensate for the inter-session variability when applied to i-vector based speaker recognition systems on short utterances. A detailed comparison of these methods with cosine similarity scoring, short utterance variance modeling, source-normalized LDA, etc. is made in Kanagasundaram et al. (2014a). It also provides proper explanation for the working of these methods. Kanagasundaram et al. (2014b) then proposes a new method that uses PLDA to model the source variance (SUV) directly. Prior to this a combination of SUVN, LDA, and SN-LDA is applied.

3.1.3 Domain Adaptation

A lot of work on domain adaptation when applied to i-vector based speaker recognition system has recently been reported. The problem of data mismatch arises when the source of test data is not known and hence the training set holds less similar feature vector values as compared to the case when training and test data are similar in nature. The performance degradation due to this data mismatch is quite significant.

Glembek et al. (2014) proposed a solution to this problem by adapting the LDA matrix to compensate for the new dataset shift. This unsupervised adaptation is done using within class covariance (WCC) for LDA which involves calculating the between-dataset low-rank covariance matrix.

Aronowitz and Rendel (2014) tried to solve this problem for the text-independent speaker verification where dot product scoring was used, which was followed by ZT-score normalization. This work tried to replace the target data with standard databases like TIMIT[1], Switchboard[2], and NIST. Hence, the target model is not adapted for any specific set of utterances. Irrelevant phonetic content was filtered and information that was usually discarded in residual supervectors was used to enhance the performance of the system. Using all the three databases together to train the target model increased the performance of the system by about 31%, which looks very promising if the data from target domain is not available apriori.

Aronowitz (2014) proposed a more generic solution to this problem by introducing an inter-dataset variability compensation (IDVC) technique to compensate for the mismatch in dataset when utterances in target and test data are from different domains. The technique tries to estimate the dataset shift vectors in i-vector domain for the training data. These i-vectors are then used to estimate a low-dimensional subspace which is then removed from test set i-vector prior to the application of PLDA. Equal error rate (EER) was reduced by 54% when this method was used as compared to the standard i-vector-PLDA system.

[1] https://catalog.ldc.upenn.edu/LDC93S1
[2] https://catalog.ldc.upenn.edu/LDC97S62

3.1.4 Normalization and Scoring

Normalization techniques are used to compensate for the noise (mismatch) in utterances due to differences in the environment. To achieve this, the feature vector is modified by scaling or warping so that the modeling of speaker differences can be done properly. Some popular techniques are cepstral mean normalization (CMN), mean and variance normalization (MVN), short-time Gaussianization (STG), short-time mean and variance normalization (STMVN), etc.

Alam et al. (2011) introduced a new method *short-time mean and scale normalization* (STMSN) and compared the performance of STG, STMVN and STMSN when used in a PLDA-based i-vector based speaker verification system with telephone and microphone data. The results showed that all the three methods perform nearly equally but STG took more time for normalization owing to its higher complexity.

Source normalization is also an important aspect to be considered while developing a speech recognition application as training set cannot possibly contain utterances collected from all possible speech sources like telephone, microphone, mobile device, etc. Hence variations are also embedded into the signals which are specific to the source. Source normalized LDA ((McLaren and van Leeuwen, 2011a,b, 2012b) improvises the LDA approach to compensate for this kind of variability. A similar approach is proposed by McLaren and van Leeuwen (2012a) for gender independent speaker recognition where modifications of WCCN are used instead of LDA to normalize the variability.

Cosine similarity is the most commonly used scoring technique employed with i-vectors since it was first introduced in Dehak (2009). It gives very good performance with an easy to implement technique but a major overhead is score normalization. Dehak et al. (2010) argued that the use of normalization can be replaced with mean and covariance which are extracted from i-vectors generated from impostor utterances. This technique gives better performance than the standard cosine similarity scoring used with ZT-norm. This performance is further improved by adapting the models defined in the total variability space in an unsupervised manner.

Bousquet et al. (2011) proposed the use of radial-NAP (nuisance attribute projection) followed by Mahalanobis metric scoring as inter-session compensation and scoring method to be used with i-vectors. This approach gives better performance as compared to LDA + WCCN + Cosine scoring.

Background normalized (B-norm) l_2 residual was used for scoring by Li et al. (2011). In this method a kind of T-norm is applied on target score by using scores generated from background score. Though l_2 residual gives inferior performance as compared to l_1 norm, its performance increases significantly when used with B-norm.

It should be noted here that very few methods other than cosine similarity are used in practice with i-vectors due to its superior performance. Few alternatives available that provide improvements are found to be more complex and hence are not very popular.

3.1.5 Data Selection

Other than normalization and scoring techniques, data selection also plays a vital role in the performance of speaker verification systems using i-vectors. Biswas et al. (2014) reduced the amount of training data by applying k-nearest neighbor (k-NN) algorithm. A basic optimization is required in this method to select k. To overcome this difficulty they also proposed a flexible k-NN (fk-NN) method which uses local distance-based outlier factor (LDOF) to select k.

3.1.6 Noisy data

One more major problem faced during speaker recognition is the noise of the ambient environment. Most of the speech recognition systems work very well on clean utterances free from any background noise, but perform quite poorly when noise is introduced into the recordings. For most of real applications, noise free data cannot be guaranteed, hence developing robust speech recognition systems which works well in the presence of noise is important.

Mandasari et al. (2012) compared the robustness of some standard speaker recognition approaches in the presence of noise. The compared techniques were GMM based dot scoring system, i-vector based LDA + WCCN + Cosine scoring system and i-vector-PLDA based system with WCCN normalization. To introduce noise into the spoken utterances NOISEX-92 (Varga and Steeneken, 1993) database was used with varying signal-to-noise (SNR) ratio. i-Vector-PLDA based system gave the least EER among the three methods. Weiner filtering was also tried to be used with i-vector-PLDA, but it did not provide sufficient performance improvement to justify its use.

Results demonstrating efficiency of i-vector-PLDA technique were also presented in Lei et al. (2012a). Garcia-Romero et al. (2012) presents a more complex version of this strategy where multiple G-PLDA subsystems are used and their results are combined using a convex mixture of scores generated by each subsystem.

Various other methods aimed at solving this problem include using Vector Taylor Series (VTS) approximation (Lei et al., 2013), multiple Support Vector Machines (SVM) which are trained using adaptive boosting (Sarkar and Rao, 2014), simplified VTS (sVTS) which reduces computational complexity of VTS (Lei et al., 2014a), acoustic feature uncertainty propagation (Yu et al., 2014), unscented transform which is used instead of VTS (Martinez et al., 2014), etc.

3.1.7 Discriminative Training

An alternate approach is to train the system to differentiate between the utterances by same speakers and different speakers. During this process the i-vectors are not modeled and this technique is referred to as discriminative training. Burget et al. (2011) used this concept to run speaker recognition on telephone speech and achieved up to 40% relative performance improvement compared to the standard method which uses generatively trained PLDA model.

Karafiát et al. (2011) presented another approach which used region dependent linear transforms (RDLT) discriminative training. RDLTs are calculated from i-vectors which consume less time because the posterior probabilities are sparse.

Cumani et al. (2012) used best second order approximation of the log-likelihood score with pairwise discriminative training for gender independent speaker recognition. The performance of the resulting system was found only a bit lower than the gender dependent system using i-vectors.

Though discriminative training gives better performance for some cases, a major disadvantage of this technique is the increased training time which is not feasible for many applications.

3.1.8 Emotional Speaker Recognition

Working on the lines of JFA and front end factor analysis, emotional variability was introduced to solve the problem of emotional speaker recognition by Chen and Yang (2011). This work describes emotional factor analysis (EFA) similar to JFA to model the emotional variability of speaker. This kind of modeling is argued to be applicable because of similarities in channel and emotional variability. i-Vector based WCCN and LDA systems were compared with EFA and GMM-UBM techniques. EFA gave the best performance in terms of low EER and higher identification rate (IR) for Mandarin affective speech corpus (MASC) (Wu et al., 2006).

The performance was further improved by compensating the emotional variability using an emotional synthesis algorithm proposed in Chen and Yang (2013). This algorithm named atom aligned sparse representation (AASR) uses the property that for each neutral utterance there is an aligned emotional utterance. The superiority of this approach was presented in terms of higher IR as compared to standard i-vector approach with LDA compensation.

3.2 i-vector in Language Recognition

Language recognition is another set of classification problem where i-vectors have been used extensively. Most of the approaches to solve this problem have used some extra features in addition to i-vectors.

After the success of i-vectors in speaker recognition, they have been used for language recognition by Martínez et al. (2011) and Dehak et al. (2011c). Here the language models are not adapted, instead the i-vectors are directly fed to any simple classifier which gives the output based on language models. The language models are generated using unsupervised training. Hence, there is no tagging of training utterances with language or speaker identity. Also i-vectors does not distinguish between inter-class and intra-session variabilities and hence contain information about both channel and class.

It is expected that like speaker recognition, the models will form clusters based on languages. Hence, it is very important that sufficient speaker variability is maintained while generating the training data. Otherwise, the models might start forming clusters based on speaker features. This problem is more dominant with i-vectors because the training is unsupervised, hence the system does not know which utterances belong to the same class.

Li and Liu (2014) used tandem features along with MFCCs to generate the extended feature vectors prior to calculating the i-vectors for the speaker and language recognition task. In this work, they manipulated the phonetic tokens and features for calculating zero and first order statistics but kept the factor analysis and i-vector calculation intact. They showed that for calculating first order statistics, tandem features are better than MFCCs in case of language identification task while the reverse holds true for speaker verification task. This work is important in the way that it showed which kind of features and tokens give better performance while applying i-vectors to various kind of recognition tasks.

3.2.1 Short Duration utterances

i-Vectors are also found to be efficient in overcoming the challenges of language recognition like short duration of utterances and less training data. i-Vector approach rely on the fact that different languages will form fairly

compact clusters due to low intra-class variability, but if the utterances are short, this variability tends to decrease the separability of clusters.Travadi et al. (2014) proposed to solve this problem by modifying the normal prior distribution of the i-vectors. It is assumed that the number of different classes is same as the number of Gaussian mixtures in prior distribution. A new parameter λ is introduced to provide more flexibility while balancing the observed data and imposed prior impact. In the i-vector estimate, the relative weights of the terms are controlled by λ.

Similar problem is solved by Segbroeck et al. (2014a) through UBM fusion. The method involves multiple UBMs, and a new supervector is generated by combining the first order statistics of all UBMs followed by training of system using EM algorithm. This method performs better in case of diverse feature representations and short utterances.

3.3 i-vector in Accent Recognition

Research has been conducted to find the accent of an utterance from spoken speech in automated systems. Recognizing accent is generally considered more difficult task as compared to language recognition owing to very little feature variations across the dialects and accents for same language (Chen et al., 2010).

Bahari et al. (2013) applied i-vectors to accent recognition and compared its performance with Gaussian mean supervector (GMS) and Gaussian posterior probability supervector (GPPS) approaches. It was further investigated that which kind of classifier among support vector machine (SVM), the Naive Bayesian classifier (NBC) and the sparse representation classifier (SRC) will give better performance with which kind of utterance modeling approach and the results showed that i-vectors are effective with the SVM classifiers to tackle the problem of accent recognition. On similar lines DeMarco and Cox (2012) classified 14 British English accents with 68% classification accuracy.

DeMarco and Cox (2013) also performed the similar analysis with native British accents and concluded that length normalization is indeed beneficial when applied along with i-vectors. This work also established that unlike speaker and language recognition tasks, neighborhood component analysis (NCA) followed by 1-NN classification is not suitable for accent recognition.

Behravan et al. (2015) argued that such high accuracy of i-vector approach might be because of the large training data available for English. Hence based on their previous work Behravan et al. (2013) which showed that i-vector approach performs better than the classic GMM-UBM approach for recognizing foreign accent from spoken Finnish, they studied various factors affecting the performance of i-vectors in this problem. The investigations were done on how the i-vector dimensionality and UBM size affect the accuracy. Also the application of Heteroscedastic LDA (HLDA) in dimensionality reduction with supervision was discussed. The language aspects like accent detection difficulty, confusion patterns between accents and effect of proficiency, education, age and knowledge of other language were also studied. The results showed that the best accuracy was achieved with 1000 dimensional i-vectors which was slightly higher than expected. The choice of UBM training data also made the largest effect on accuracy.

3.4 i-vector in AED

Acoustic event detection (AED) refers to the task of detecting any acoustic event from an audio or an utterance. Many kinds of such events other than speech (most informative of all) like clapping, laughter, yawning, etc. may be helpful in analyzing the environment features where audio is generated. These events generally arise due to sounds generated by the human body or by external factors. Hence detecting these events may help in describing the person or identifying the social activity taking place in the surrounding environment.

In the field of AED, i-vectors have been used with blind segmentation approach by Huang et al. (2013) to achieve 8% better F_1 score than classic HMM based systems. Here the segmentation is done randomly by dividing streams into pieces of equal length. Then multiple categories are labeled on acoustic events and boundary information is not stored. i-Vectors are then extracted and classification is performed on them. This approach saves the issues of manual labeling and overlapping events.

3.5 i-vector in Speech Diarization

Speech diarization is the process of splitting the utterance or audio into parts such that each part correspond to a single speaker. It combines speaker recognition and speaker clustering. Hence it is important to find the points where speaker changes and then cluster the partitions of same speaker together. It finds application in creating transcripts of audio data.

Shum et al. (2011) proposed a total variability based approach for speaker diarization task. It provided state of the art results on two-speaker telephone diarization task. It was argued that compensation for inter-session variability is not required as thought previously. The

utterance is divided into segments and i-vectors were extracted for each segment. Then their dimensionality was further reduced using principal component analysis (PCA). K-means clustering with $K = 2$ was applied on these dimensionally reduced i-vectors and the results were refined using Viterbi re-segmentation and Baum–Welch soft speaker clustering algorithm. The final step involves generating a single i-vector for each speaker using new assignments and reassigning i-vector of each segment to speakers based on cosine similarity.

Dupuy et al. (2012) implemented the similar strategy for multiple-speaker diarization task. Here the number of speakers were not known and the task was to detect speakers appearing in shows. Since the value of K is not known for applying K-means, this clustering problem was expressed as an integer linear programming (ILP) problem. The results were then compared with normalized cross-likelihood ratio (NCLR) based speaker diarization system proposed by Le et al. (2007). The error rates of both methods were similar, but i-vector based ILP approach was found to be about 8.66 times faster.

Zheng et al. (2014) recently combined the variational Bayes approach with i-vectors to improve the diarization error rate (DER) for two-speaker telephone diarization task.

3.6 i-vector in Other Domains

After the success of i-vectors in several speech related domains was established, their extensions to allied domains were also investigated. One such area is emotion recognition. Mariooryad and Busso (2014) proposed the use of factor analysis for modeling the emotion recognition task from speech utterances. Interactive emotional dyadic motion capture (IEMOCAP) database (Busso et al., 2008) was used to perform the recognition tasks. Along with recognizing emotions, factorization of speaker characteristics and classification of expressive behaviors were also targeted in this work.

Xia and Liu (2012) proposed that instead of using i-vectors directly, it might be beneficial to use some concatenated i-vector features which are emotion specific. These are then used in SVM classifiers and then LDA and WCCN are employed as dimensionality reduction techniques to recognize the emotions.

Some other applications of i-vectors include gesture recognition (Cheng et al., 2014), age estimation (Bahari et al., 2014), event detection in consumer media (Zhuang et al., 2012), classification of Cognitive Load from speech (Segbroeck et al., 2014b), forensics (Mandasari et al., 2011), joint anti-spoofing (Sizov et al., 2015), etc.

4 Hybrid Approaches

4.1 Prosodic Approach

Prosodic features such as rhythm, intonation and stress play an important role in speech processing. Some of the approaches to use them for speaker verification were explored by Adami et al. (2003), Adami (2007), Ferrer et al. (2010), Kockmann et al. (2010), etc. JFA based approaches using prosody (Dehak et al., 2007a) generated features from some or all prosodic features instead of cepstral features. They are then modeled using GMMs and FA is applied to them. (Dehak et al., 2007b) further enhanced the performance by using formants with prosodic contours to generate the features.

One of the starting attempts to use prosodic features for language recognition were used in Foil (1986). Short noisy speech segments were used as a database to test the system while a formant clustering algorithm was used to reach at some decision.

4.2 Combination of Cepstral and Prosodic Approach

A combination of prosodic and cepstral features has also been used in speaker recognition in Kockmann et al. (2011). Here i-vectors were generated from cepstrals and prosodic features separately and then were concatenated to get new set of i-vectors. These were used to train a PLDA model to get superior performance as compared to traditional i-vector techniques.

i-Vector based approach where the fusion of acoustic system (cepstral features) and prosodic system to improve the performance as compared to both the systems alone was used in Martinez et al. (2012, 2013).

4.3 Phonotactic Approach

Another important language recognition paradigm in addition to acoustic and prosodic is the phonotactic approach. Since phonotactic constraints are highly language dependent, they are used successfully in LRE. In this approach, the speech utterances are tokenized into discrete events using phone recognizers. These tokens are then used to extract the n-gram counts for providing input to the classifiers. If discriminative classifiers are used then n-gram counts are represented as vectors of fixed length. The size of phoneme inventory of the language decides the dimensionality of the vector. i-Vectors solve the problem of reducing this dimensionality thereby decreasing the training and classification time (Soufifar et al., 2011).

4.3.1 i-vector with DNNs and DBNs

The use of i-vectors with deep neural networks (DNNs) showed reduction in word error rate (WER) as compared to normal i-vector implementations. Here i-vectors are appended with basic acoustic features, which are then provided as input for neural network training. Using i-vectors helps DNN in differentiating between phonetic events by taking more speaker and session variability into account. Details regarding the use of i-vectors with DNN can be found in Gupta et al. (2014), Variani et al. (2014), Lei et al. (2014b), Rouvier and Favre (2014) and Karanasou et al. (2014).

Senior and Lopez-Moreno (2014) takes up another important issue of overfitting to i-vectors by DNN, resulting in slightly higher WERs. Two solutions to this problem are proposed, viz., reducing the dimensionality of i-vectors and regularizing the network parameters. Using regularization to solve this issue resulted in reduced WERs irrespective of the model sizes.

Ghahabi and Hernando (2014a) used deep belief networks (DBNs) to model the impostor and target speaker models discriminatingly. The proposed method also balances the amount of positive and negative input data. Then a Universal DBN is trained by using many i-vectors from different background speakers. For a fewer number of input samples, adaptation is also done alongwith pre-training to get a good target model. Fine tuning is then done by adding one more label layer to the system. This technique outperformed the cosine classifier and conventional neural networks. Ghahabi and Hernando (2014b) is another work discussing the global impostor selection for DBNs by iteratively dividing the impostor i-vector database randomly in two sets and generating centroids to represent the global impostors. This system reported about 22% performance improvement over the baseline.

Some works also try to question the assumptions made by front end factor analysis. One such assumption was analyzed by Lei et al. (2012b). They relaxed the assumption that speaker identity is independent of intersession variability. Hence ,if a speaker changes location then it would effectively also change the within-speaker or intersession variability along with intra-session variability. The proposed method was termed as bilinear factor analysis for i-vectors since the relation between latent variables describing the speaker and the channel was considered to be bilinear. The method showed superior performance to i-vectors but the configuration on which i-vector performance was compared was not optimal. This was attributed to the fact that the same vector size calculation in bilinear factor analysis was not feasible as it consisted of matrix inversion for high dimensions.

5 Toolkits

Since the total variability analysis approach has been proposed, a few libraries have started providing the required methods as a part of their standard toolkit. A few of them are listed in this section.

5.1 The Kaldi Speech Recognition Toolkit [3]

C++ based Kaldi Speech Recognition Toolkit (Povey et al., 2011) was first introduced in 2011 as a speech recognition system which was based on finite-state transducers. It later provided support for i-vectors through the header file `ivector-extractor.h`. Support for linear algebra is quite extensive and works well with most of the widely available speech databases. Parallelization support is also provided with this toolkit.

One major advantage is that it supports open and distributed development model, and hence can be modified easily. The only drawback is that the documentation of this toolkit is mostly meant for the experts using this toolkit. A proper comparison of Kaldi with other open source ASR systems like HTK (Young et al., 2006), pocketSphinx (Huggins-Daines et al., 2006), and Sphinx-4 (Lamere et al., 2003) can be found in Gaida et al. (2014).

5.2 ALIZE 3.0 [4]

The ALIZE Toolkit (Larcher et al., 2013) provides a wide variety of methods for complete i-vector extraction and processing. It has a full set of methods for normalization and scoring. It is also written in C++ and provides implementations for JFA as well. It is provided under LGPL licence. ALIZE consists of two major packages:

- ALIZE package: It is a low level library that consists of all the methods related to Gaussian mixtures.
- LIA_RAL package: It is a high level package aimed at providing programs related to Automatic Speaker Detection. It consists of the following parts:
  - LIA_SpkTools: It is a high level library containing finished tools provided by LIA_RAL.
  - LIA_SpkDet: It consists of tools required for any task related to biometric authentication system.
  - LIA_Utils: It contains tools required in changing ALIZE related formats like GMM models, any parameters, etc.

[3] http://kaldi.sourceforge.net
[4] http://alize.univ-avignon.fr

– LIA_SpkSeg: It is a recent tool added to ALIZE toolkit which is to be used for Speaker Diarization.

ALIZE is one of the most complete set of libraries that are required to develop a speech recognition application. Other than GMM-UBM model, this toolkit also consists of implementation for JFA and i-vectors. The wide range of scoring methods inbuilt in the system are also sufficient for most practical systems. Though ALIZE provides nearly all the methods required for performing any kind of speech recognition, it is difficult to start with due to insufficient documentation provided with the toolkit.

5.3 MSR Identity Toolbox [5]

MSR Identity Toolbox (Sadjadi et al., 2013) is a MATLAB based toolbox which supports GMM-UBM and i-vector-PLDA based approaches for speaker recognition. It aims at providing basic methods so as to enable developing the baseline systems quickly. It can also be used for other applications mentioned in this paper, viz language, accent, dialect recognition, etc.

This toolkit has some of the easiest implementations of i-vector system and can be learned easily. It consists of demos for GMM-UBM and i-vector systems with complete step-by-step explanations. The code is implemented such that it is easier for MATLAB to vectorize it. Also the code is parallelizable when used with the parallel computing toolbox.

The toolkit contains the support for the following major functions: (1) feature normalization; (2) GMM-UBM system; (3) i-Vector-PLDA, and; (4) equal error rate (EER), detection cost function (DCF) and detection error tradeoff (DET) plots.

5.4 LIUM SpkDiarization [6]

LIUM SpkDiarization (Meignier and Merlin, 2010) is a toolkit specialized for Speech Diarization application. Rouvier et al. (2013) presented its application in broadcast news diarization which uses integer linear programming (ILP) as a clustering algorithm in its new version. Similarity modeling and measurement is done using i-vectors in this toolkit. It consists of a complete set of functionalities for speaker diarization including feature extraction, silence detection and standard diarization tools. Sphinx 4 classes can be used in LIUM to read the acoustic features dynamically. LIUM also supports some standard file formats like Sphinx format, SPro4 format, HTK format and gzipped-text (in which each line corresponds to a vector). The implementation is done in Java and is similar to Sphinx. This toolkit is preferred for Radio or TV Show transcription and is not expected to give high performance for meeting or conversation transcription.

[5] http://research.microsoft.com/en-us/downloads/2476c44a-1f63-4fe0-b805-8c2de395bb2c/

[6] https://projets-lium.univ-lemans.fr/spkdiarization/ Link updated in March 2021.

6 Future of i-vectors

A lot of progress in factor analysis techniques for speech recognition applications have been achieved by the research community in the last decade. This effort has led to the development of i-vectors which has become the state of the art technique in a very short period of time. The amount of work done in its use for applications other than speaker recognition is overwhelming.

These methods are effective but still suffer from some bottlenecks as explained earlier. One major problem is the data selection dependency of the i-vector training. If the data is not properly chosen for enrollment set the recognition system might perform poorly when run with a test set. Sometimes it becomes necessary to generate the gender dependent PLDA models to boost the recognition performance. For using the recognition systems practically this kind of dependence must be tackled, otherwise the amount of data on which the models are trained should be kept very high so that maximum number of characteristics are modeled in the target models.

The current approaches also give low performance for short utterances (Verma, 2015). Though many methods have been proposed to tackle this issue, very few can be used practically as most of them put strong assumptions on the data on which the system is trained and tested. The problem becomes more intriguing because most of the methods developed are for long utterances and does not work as it is on small utterances. Sufficient changes should be made in their design to get acceptable performance out of the system. This problem should be the main focus for future approaches as most of the speech recognition applications need to be run with very short utterances.

7 Conclusion

We have presented a concise overview of various speech recognition methods that uses i-vectors. The recognition performance of these methods vary across the domains. Some of them are highly domain or problem specific

and generally perform well in ideal laboratory environments. The main challenges are found to be the paucity of training data, choice of training data, length of utterances, background/environmental noise, emotional state of speaker, etc. Hence most of the methods fail when tested in real life environments. Though some of the methods looks highly promising, much work needs to be done in coming up with a generic method that is suitable to tackle all the challenges mentioned. Another important criteria for use of i-vectors is the execution time of the recognition process. Though some methods might give very good performance, if the running time on real data is high, they may render the application useless for most practical situations.

Recently, NIST has started promoting the research based on i-vectors. It has conducted an *i-vector Machine Learning Challenge*[7] for speaker recognition evaluation (SRE) task in 2013-14. The same challenge for language recognition evaluation (LRE) task has been started in 2015. Many toolkits are also now avaiable that has eased the deployment of i-vectors. We hope that these initiatives will foster the use of i-vectors in diverse domains.

Acknowledgements

The authors wish to acknowledge UNICEF India, and the DST, Government of India, for the funding provided under their FIST scheme, which greatly aided in the work reported herein.

References

Adami A, Mihaescu R, Reynolds D, Godfrey J (2003) Modeling prosodic dynamics for speaker recognition. In: Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on, vol 4, pp IV–788–91 vol.4, DOI 10.1109/ICASSP.2003.1202761

Adami AG (2007) Modeling prosodic differences for speaker recognition. Speech Commun 49(4):277–291, DOI 10.1016/j.specom.2007.02.005, URL `http://dx.doi.org/10.1016/j.specom.2007.02.005`

Alam MJ, Ouellet P, Kenny P, O'Shaughnessy DD (2011) Comparative evaluation of feature normalization techniques for speaker verification. In: Advances in Nonlinear Speech Processing - 5th International Conference on Nonlinear Speech Processing, NOLISP 2011, Las Palmas de Gran Canaria, Spain, November 7-9, 2011. Proceedings, pp 246–253, DOI 10.1007/978-3-642-25020-0_32

Aronowitz H (2014) Inter dataset variability compensation for speaker recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp 4002–4006, DOI 10.1109/ICASSP.2014.6854353

Aronowitz H, Barkan O (2012) Efficient approximated i-vector extraction. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp 4789–4792, DOI 10.1109/ICASSP.2012.6288990

Aronowitz H, Rendel A (2014) Domain adaptation for text dependent speaker verification. In: INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014, pp 1337–1341

Bahari M, Saeidi R, Van hamme H, Van Leeuwen D (2013) Accent recognition using i-vector, gaussian mean supervector and gaussian posterior probability supervector for spontaneous telephone speech. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp 7344–7348, DOI 10.1109/ICASSP.2013.6639089

Bahari MH, McLaren M, hamme HV, van Leeuwen DA (2014) Speaker age estimation using i-vectors. Eng Appl of AI 34:99–108, DOI 10.1016/j.engappai.2014.05.003, URL `http://dx.doi.org/10.1016/j.engappai.2014.05.003`

Behravan H, Hautamäki V, Kinnunen T (2013) Foreign accent detection from spoken finnish using i-vectors. In: INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013, pp 79–83

Behravan H, Hautamäki V, Kinnunen T (2015) Factors affecting i-vector based foreign accent recognition: A case study in spoken finnish. Speech Communication 66(0):118 – 129, DOI http://dx.doi.org/10.1016/j.specom.2014.10.004

Biswas S, Rohdin J, Shinoda K (2014) i-Vector Selection for Effective PLDA Modeling in Speaker Recognition. In: Proc. Odyssey 2014 - The Speaker and Language Recognition Workshop, pp 100–105

Bousquet P, Matrouf D, Bonastre J (2011) Intersession compensation and scoring methods in the i-vectors space for speaker recognition. In: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011, pp 485–488

Brümmer N, Strasheim A, Hubeika V, Matejka P, Burget L, Glembek O (2009) Discriminative acoustic language recognition via channel-compensated GMM statistics. In: INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association,

[7] https://www.nist.gov/itl/iad/mig/i-vector-machine-learning-challenge. Link updated in March 2021.

Brighton, United Kingdom, September 6-10, 2009, pp 2187–2190

Burget L, Plchot O, Cumani S, Glembek O, Matejka P, Brümmer N (2011) Discriminatively trained probabilistic linear discriminant analysis for speaker verification. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, May 22-27, 2011, Prague Congress Center, Prague, Czech Republic, pp 4832–4835, DOI 10.1109/ICASSP.2011.5947437

Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang J, Lee S, Narayanan S (2008) Iemocap: interactive emotional dyadic motion capture database. Language Resources and Evaluation 42(4):335–359, DOI 10.1007/s10579-008-9076-6, URL http://dx.doi.org/10.1007/s10579-008-9076-6

Chen L, Yang Y (2011) Applying emotional factor analysis and i-vector to emotional speaker recognition. In: Sun Z, Lai J, Chen X, Tan T (eds) Biometric Recognition, Lecture Notes in Computer Science, vol 7098, Springer Berlin Heidelberg, pp 174–179, DOI 10.1007/978-3-642-25449-9_22, URL http://dx.doi.org/10.1007/978-3-642-25449-9_22

Chen L, Yang Y (2013) Emotional speaker recognition based on i-vector through atom aligned sparse representation. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp 7760–7764, DOI 10.1109/ICASSP.2013.6639174

Chen N, Shen W, Campbell J (2010) A linguistically-informative approach to dialect recognition using dialect-discriminating context-dependent phonetic models. In: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, pp 5014–5017, DOI 10.1109/ICASSP.2010.5495068

Cheng YC, Hautamaki V, Huang Z, Li K, Lee CH (2014) An i-vector based descriptor for alphabetical gesture recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp 6593–6597, DOI 10.1109/ICASSP.2014.6854875

Cumani S, Glembek O, Brümmer N, de Villiers E, Laface P (2012) Gender independent discriminative speaker recognition in i-vector space. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012, Kyoto, Japan, March 25-30, 2012, pp 4361–4364, DOI 10.1109/ICASSP.2012.6288885

Dehak N (2009) Discriminative and Generative Approaches for Long- and Short-term Speaker Characteristics Modeling: Application to Speaker Verification. PhD thesis, Ecole de Technologie Superieure (Canada), aAINR50490

Dehak N, Shum S (2011) Low-dimensional speech representation based on factor analysis and its applications. Johns Hopkins CLSP Lecture

Dehak N, Dumouchel P, Kenny P (2007a) Modeling prosodic features with joint factor analysis for speaker verification. Audio, Speech, and Language Processing, IEEE Transactions on 15(7):2095–2103, DOI 10.1109/TASL.2007.902758

Dehak N, Kenny P, Dumouchel P (2007b) Continuous prosodic features and formant modeling with joint factor analysis for speaker verification. In: INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007, pp 1234–1237

Dehak N, Dehak R, Glass JR, Reynolds DA, Kenny P (2010) Cosine similarity scoring without score normalization techniques. In: Proc. Odyssey 2010 - The Speaker and Language Recognition Workshop, pp 15–19

Dehak N, Karam ZN, Reynolds DA, Dehak R, Campbell WM, Glass JR (2011a) A channel-blind system for speaker verification. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, May 22-27, 2011, Prague Congress Center, Prague, Czech Republic, pp 4536–4539, DOI 10.1109/ICASSP.2011.5947363

Dehak N, Kenny P, Dehak R, Dumouchel P, Ouellet P (2011b) Front-end factor analysis for speaker verification. Audio, Speech, and Language Processing, IEEE Transactions on 19(4):788–798, DOI 10.1109/TASL.2010.2064307

Dehak N, Torres-Carrasquillo PA, Reynolds DA, Dehak R (2011c) Language recognition via i-vectors and dimensionality reduction. In: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011, pp 857–860

DeMarco A, Cox SJ (2012) Iterative classification of regional British accents in i-vector space. In: 2012 Symposium on Machine Learning in Speech and Language Processing, MLSLP 2012, Portland, Oregon, USA, September 14, 2012, pp 1–4

DeMarco A, Cox SJ (2013) Native accent classification via i-vectors and speaker compensation fusion. In: INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013, pp 1472–1476

Dupuy G, Rouvier M, Meignier S, Estève Y (2012) I-vectors and ILP clustering adapted to cross-show speaker diarization. In: INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012, pp 2174–2177

Ferrer L, Scheffer N, Shriberg E (2010) A comparison of approaches for modeling prosodic features in speaker recognition. In: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, pp 4414–4417, DOI 10.1109/ICASSP.2010.5495632

Foil J (1986) Language identification using noisy speech. In: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86., vol 11, pp 861–864, DOI 10.1109/ICASSP.1986.1168879

Gaida C, Lange P, Petrick R, Proba P, Malatawy A, Suendermann-Oeft D (2014) Comparing open-source speech recognition toolkits. URL `http://suendermann.com/su/pdf/oasis2014.pdf`

Garcia-Romero D, Espy-Wilson CY (2011) Analysis of i-vector length normalization in speaker recognition systems. In: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011, pp 249–252

Garcia-Romero D, Zhou X, Espy-Wilson CY (2012) Multicondition training of gaussian plda models in i-vector space for noise and reverberation robust speaker recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp 4257–4260, DOI 10.1109/ICASSP.2012.6288859

Ghahabi O, Hernando J (2014a) Deep belief networks for i-vector based speaker recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014, pp 1700–1704, DOI 10.1109/ICASSP.2014.6853888

Ghahabi O, Hernando J (2014b) Global impostor selection for dbns in multi-session i-vector speaker recognition. In: Advances in Speech and Language Technologies for Iberian Languages - Second International Conference, IberSPEECH 2014, Las Palmas de Gran Canaria, Spain, November 19-21, 2014. Proceedings, pp 89–98, DOI 10.1007/978-3-319-13623-3

Glembek O, Burget L, Dehak N, Brummer N, Kenny P (2009) 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, DOI 10.1109/ICASSP.2009.4960519

Glembek O, Burget L, Matejka P, Karafiat M, Kenny P (2011) Simplification and optimization of i-vector extraction. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp 4516–4519, DOI 10.1109/ICASSP.2011.5947358

Glembek O, Ma J, Matejka P, Zhang B, Plchot O, Burget L, Matsoukas S (2014) Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp 4032–4036, DOI 10.1109/ICASSP.2014.6854359

Gupta V, Kenny P, Ouellet P, Stafylakis T (2014) I-vector-based speaker adaptation of deep neural networks for french broadcast audio transcription. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp 6334–6338, DOI 10.1109/ICASSP.2014.6854823

Hasan T, Saeidi R, Hansen J, van Leeuwen D (2013) Duration mismatch compensation for i-vector based speaker recognition systems. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp 7663–7667, DOI 10.1109/ICASSP.2013.6639154

Hautamäki V, Cheng Y, Rajan P, Lee C (2013) Minimax i-vector extractor for short duration speaker verification. In: INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013, pp 3708–3712

Huang Z, Cheng Y, Li K, Hautamäki V, Lee C (2013) A blind segmentation approach to acoustic event detection based on i-vector. In: INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013, pp 2282–2286

Huggins-Daines D, Kumar M, Chan A, Black A, Ravishankar M, Rudnicky A (2006) Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on, vol 1, pp I–I, DOI 10.1109/ICASSP.2006.1659988

Jancik Z, Plchot O, Brummer N, Burget L, Glembek O, Hubeika V, Karafiat M, Matejka P, Mikolov T, Strasheim A, Cernocky J (2010) Data selection and calibration issues in automatic language recognition - investigation with but-agnitio nist lre 2009 system. In: Proc. Odyssey 2010 - The Speaker and Language Recognition Workshop, pp 215–221

Jiang Y, Lee K, Tang Z, Ma B, Larcher A, Li H (2012) PLDA modeling in i-vector and supervector space for speaker verification. In: INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012, pp 1680–1683

Kanagasundaram A, Vogt R, Dean D, Sridharan S, Mason M (2011) i-vector based speaker recognition on short utterances. In: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31,

2011, pp 2341–2344

Kanagasundaram A, Dean D, Vogt R, McLaren M, Subramanian S, Mason M (2012a) Weighted LDA techniques for i-vector based speaker verification. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012, Kyoto, Japan, March 25-30, 2012, pp 4781–4784, DOI 10.1109/ICASSP.2012.6288988

Kanagasundaram A, Vogt R, Dean D, Sridharan S (2012b) PLDA based speaker recognition on short utterances. In: Odyssey 2012: The Speaker and Language Recognition Workshop, Singapore, June 25-28, 2012, pp 28–33

Kanagasundaram A, Dean D, Gonzalez-Dominguez J, Sridharan S, Ramos D, Gonzalez-Rodriguez J (2013) Improving short utterance based i-vector speaker recognition using source and utterance-duration normalization techniques. In: INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013, pp 2465–2469

Kanagasundaram A, Dean D, Sridharan S, Gonzalez-Dominguez J, Gonzalez-Rodriguez J, Ramos D (2014a) Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques. Speech Communication 59(0):69 – 82, DOI http://dx.doi.org/10.1016/j.specom.2014.01.004

Kanagasundaram A, Dean D, Sridharan S, Gonzalez-Dominguez J, González-Rodríguez J, Ramos D (2014b) Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques. Speech Communication 59:69–82, DOI 10.1016/j.specom.2014.01.004

Karafiát M, Burget L, Matejka P, Glembek O, Cernocký J (2011) ivector-based discriminative adaptation for automatic speech recognition. In: 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2011, Waikoloa, HI, USA, December 11-15, 2011, pp 152–157, DOI 10.1109/ASRU.2011.6163922

Karanasou P, Wang Y, Gales MJF, Woodland PC (2014) Adaptation of deep neural network acoustic models using factorised i-vectors. In: INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014, pp 2180–2184

Kenny P (2005) Joint factor analysis of speaker and session variability: Theory and algorithms. Tech. Rep. CRIM-06/08-13, Centre de Recherche Informatique de Montreal (CRIM)

Kenny P, Boulianne G, Ouellet P, Dumouchel P (2007a) Joint factor analysis versus eigenchannels in speaker recognition. Audio, Speech, and Language Processing, IEEE Transactions on 15(4):1435–1447, DOI 10.1109/TASL.2006.881693

Kenny P, Boulianne G, Ouellet P, Dumouchel P (2007b) Speaker and session variability in gmm-based speaker verification. Audio, Speech, and Language Processing, IEEE Transactions on 15(4):1448–1460, DOI 10.1109/TASL.2007.894527

Kenny P, Ouellet P, Dehak N, Gupta V, Dumouchel P (2008) A study of interspeaker variability in speaker verification. Audio, Speech, and Language Processing, IEEE Transactions on 16(5):980–988, DOI 10.1109/TASL.2008.925147

Kenny P, Stafylakis T, Ouellet P, Alam M, Dumouchel P (2013) Plda for speaker verification with utterances of arbitrary duration. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp 7649–7653, DOI 10.1109/ICASSP.2013.6639151

Kockmann M, Burget L, Cernocký J (2010) Brno university of technology system for interspeech 2010 paralinguistic challenge. In: INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010, pp 2822–2825

Kockmann M, Ferrer L, Burget L, Cernocký J (2011) ivector fusion of prosodic and cepstral features for speaker verification. In: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011, pp 265–268

Lamere P, Kwok P, Gouvea E, Raj B, Singh R, Walker W, Warmuth M, Wolf P (2003) The CMU SPHINX-4 speech recognition system. IEEE Intl Conf on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong 1:2–5

Larcher A, Bousquet P, Lee KA, Matrouf D, Li H, Bonastre JF (2012) I-vectors in the context of phonetically-constrained short utterances for speaker verification. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp 4773–4776, DOI 10.1109/ICASSP.2012.6288986

Larcher A, Bonastre J, Fauve BGB, Lee K, Lévy C, Li H, Mason JSD, Parfait J (2013) ALIZE 3.0 - open source toolkit for state-of-the-art speaker recognition. In: INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013, pp 2768–2772

Le VB, Mella O, Fohr D (2007) Speaker diarization using normalized cross likelihood ratio. In: INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007, pp 1869–1872

Lei Y, Burget L, Ferrer L, Graciarena M, Scheffer N (2012a) Towards noise-robust speaker recognition using probabilistic linear discriminant analysis. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, IEEE, pp 4253–4256

Lei Y, Burget L, Scheffer N (2012b) Bilinear factor analysis for ivector based speaker verification. In: INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012, pp 1588–1591

Lei Y, Burget L, Scheffer N (2013) A noise robust i-vector extractor using vector taylor series for speaker recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp 6788–6791, DOI 10.1109/ICASSP.2013.6638976

Lei Y, McLaren M, Ferrer L, Scheffer N (2014a) Simplified vts-based i-vector extraction in noise-robust speaker recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp 4037–4041, DOI 10.1109/ICASSP.2014.6854360

Lei Y, Scheffer N, Ferrer L, McLaren M (2014b) A novel scheme for speaker recognition using a phonetically-aware deep neural network. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp 1695–1699, DOI 10.1109/ICASSP.2014.6853887

Li M, Liu W (2014) Speaker verification and spoken language identification using a generalized i-vector framework with phonetic tokenizations and tandem features. In: INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014, pp 1120–1124

Li M, Zhang X, Yan Y, Narayanan SS (2011) Speaker verification using sparse representations on total variability i-vectors. In: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011, pp 2729–2732

Mandasari M, McLaren M, van Leeuwen D (2012) The effect of noise on modern automatic speaker recognition systems. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp 4249–4252, DOI 10.1109/ICASSP.2012.6288857

Mandasari MI, McLaren M, van Leeuwen DA (2011) Evaluation of i-vector speaker recognition systems for forensic application. In: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011, pp 21–24

Mariooryad S, Busso C (2014) Compensating for speaker or lexical variabilities in speech for emotion recognition. Speech Communication 57(0):1 – 12, DOI http://dx.doi.org/10.1016/j.specom.2013.07.011, URL `http://www.sciencedirect.com/science/article/pii/S0167639313001015`

Martínez D, Plchot O, Burget L, Glembek O, Matejka P (2011) Language recognition in ivectors space. In: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011, pp 861–864

Martinez D, Burget L, Ferrer L, Scheffer N (2012) ivector-based prosodic system for language identification. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp 4861–4864, DOI 10.1109/ICASSP.2012.6289008

Martinez D, Lleida E, Ortega A, Miguel A (2013) Prosodic features and formant modeling for an ivector-based language recognition system. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp 6847–6851, DOI 10.1109/ICASSP.2013.6638988

Martinez D, Burget L, Stafylakis T, Lei Y, Kenny P, Lleida E (2014) Unscented transform for ivector-based noisy speaker recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp 4042–4046, DOI 10.1109/ICASSP.2014.6854361

Matejka P, Glembek O, Castaldo F, Alam M, Plchot O, Kenny P, Burget L, Cernocky J (2011) Full-covariance ubm and heavy-tailed plda in i-vector speaker verification. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp 4828–4831, DOI 10.1109/ICASSP.2011.5947436

McLaren M, van Leeuwen D (2011a) Source-normalised-and-weighted lda for robust speaker recognition using i-vectors. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp 5456–5459, DOI 10.1109/ICASSP.2011.5947593

McLaren M, van Leeuwen D (2012a) Gender-independent speaker recognition using source normalisation. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp 4373–4376, DOI 10.1109/ICASSP.2012.6288888

McLaren M, van Leeuwen D (2012b) Source-normalized lda for robust speaker recognition using i-vectors from multiple speech sources. Audio, Speech, and Language Processing, IEEE Transactions on 20(3):755–766, DOI 10.1109/TASL.2011.2164533

McLaren M, van Leeuwen DA (2011b) To weight or not to weight: Source-normalised LDA for speaker recognition using i-vectors. In: INTERSPEECH 2011,

12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011, pp 2709–2712

Meignier S, Merlin T (2010) LIUM SpkDiarization: an open source toolkit for diarization. In: in CMU SPUD Workshop, vol 2010

Novoselov S, Pekhovsky T, Simonchik K, Shulipa A (2014) RBM-PLDA subsystem for the NIST i-vector challenge. In: INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014, pp 378–382

Picone JW (1993) Signal modeling techniques in speech recognition. Proceedings of the IEEE 81(9):1215–1247, DOI 10.1109/5.237532

Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Schwarz P, Silovsky J, Stemmer G, Vesely K (2011) The Kaldi Speech Recognition Toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, IEEE Signal Processing Society, iEEE Catalog No.: CFP11SRW-USB

Reynolds DA (1992) A Gaussian mixture modeling approach to text-independent speaker identification. Ph.D. dissertation, Georgia Institute of Technology

Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted gaussian mixture models. Digital Signal Processing 10(1–3):19 – 41, DOI http://dx.doi.org/10.1006/dspr.1999.0361

Rouvier M, Favre B (2014) Speaker adaptation of dnn-based ASR with i-vectors: does it actually adapt models to speakers? In: INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014, pp 3007–3011

Rouvier M, Dupuy G, Gay P, el Khoury E, Merlin T, Meignier S (2013) An open-source state-of-the-art toolbox for broadcast news diarization. In: INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013, pp 1477–1481

Sadjadi SO, Slaney M, Heck L (2013) Msr identity toolbox v1.0: A matlab toolbox for speaker recognition research. Tech. Rep. MSR-TR-2013-133

Sarkar AK, Matrouf D, Bousquet P, Bonastre J (2012) Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification. In: INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012, pp 2662–2665

Sarkar S, Rao KS (2014) A novel boosting algorithm for improved i-vector based speaker verification in noisy environments. In: INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014, pp 671–675

Segbroeck MV, Travadi R, Narayanan SS (2014a) UBM fused total variability modeling for language identification. In: INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014, pp 3027–3031

Segbroeck MV, Travadi R, Vaz C, Kim J, Black MP, Potamianos A, Narayanan SS (2014b) Classification of cognitive load from speech using an i-vector framework. In: INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014, pp 751–755

Senior A, Lopez-Moreno I (2014) Improving DNN speaker independence with i-vector inputs. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp 225–229, DOI 10.1109/ICASSP.2014.6853591

Senoussaoui M, Kenny P, Dehak N, Dumouchel P (2010) An i-vector extractor suitable for speaker recognition with both microphone and telephone speech. In: Odyssey 2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, June 28 - July 1, 2010, p 6

Senoussaoui M, Kenny P, Brümmer N, de Villiers E, Dumouchel P (2011) Mixture of PLDA models in i-vector space for gender-independent speaker recognition. In: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011, pp 25–28

Shum S, Dehak N, Chuangsuwanich E, Reynolds DA, Glass JR (2011) Exploiting intra-conversation variability for speaker diarization. In: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011, pp 945–948

Silovsky J, Prazak J (2012) Speaker diarization of broadcast streams using two-stage clustering based on i-vectors and cosine distance scoring. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp 4193–4196, DOI 10.1109/ICASSP.2012.6288843

Simonchik K, Pekhovsky T, Shulipa A, Afanasyev A (2012) Supervized mixture of PLDA models for cross-channel speaker verification. In: INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012, pp 1684–1687

Sizov A, el Khoury E, Kinnunen T, Wu Z, Marcel S (2015) Joint speaker verification and antispoofing in the i-vector space. IEEE Transactions on Information Forensics and Security 10(4):821–832, DOI 10.1109/TIFS.2015.2407362, URL http://dx.doi.org/10.1109/TIFS.2015.2407362

Soufifar M, Kockmann M, Burget L, Plchot O, Glembek O, Svendsen T (2011) ivector approach to phonotactic language recognition. In: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011, pp 2913–2916

Travadi R, Segbroeck MV, Narayanan SS (2014) Modified-prior i-vector estimation for language identification of short duration utterances. In: INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014, pp 3037–3041

Varga A, Steeneken HJ (1993) Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication 12(3):247 – 251, DOI http://dx.doi.org/10.1016/0167-6393(93)90095-3, URL http://www.sciencedirect.com/science/article/pii/0167639393900953

Variani E, Lei X, McDermott E, Lopez Moreno I, Gonzalez-Dominguez J (2014) Deep neural networks for small footprint text-dependent speaker verification. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp 4052–4056, DOI 10.1109/ICASSP.2014.6854363

Verma P (2015) Resource Usage Analysis for Speech Recognition Techniques. Master's thesis, Department of Computer Science & Engineering, Indian Institute of Technology Guwahati, India

Villalba J, Lleida E (2013) Handling i-vectors from different recording conditions using multi-channel simplified plda in speaker recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp 6763–6767, DOI 10.1109/ICASSP.2013.6638971

Wolf JJ (1972) Efficient acoustic parameters for speaker recognition. The Journal of the Acoustical Society of America 51(6B):2044–2056, DOI http://dx.doi.org/10.1121/1.1913065, URL http://scitation.aip.org/content/asa/journal/jasa/51/6B/10.1121/1.1913065

Wu T, Yang Y, Wu Z, Li D (2006) MASC: A Speech Corpus in Mandarin for Emotion Analysis and Affective Speaker Recognition. In: Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The, pp 1–5, DOI 10.1109/ODYSSEY.2006.248084

Xia R, Liu Y (2012) Using i-vector space model for emotion recognition. In: INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012, pp 2230–2233

Yin SC, Kenny P, Rose R (2006) Experiments in speaker adaptation for factor analysis based speaker verification. In: Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The, pp 1–6, DOI 10.1109/ODYSSEY.2006.248130

Young S, Evermann G, Gales M, Hain T, Kershaw D, Liu XA, Moore G, Odell J, Ollason D, Povey D, et al. (2006) The htk book (for htk version 3.4)

Yu C, Liu G, Hahm S, Hansen J (2014) Uncertainty propagation in front end factor analysis for noise robust speaker recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp 4017–4021, DOI 10.1109/ICASSP.2014.6854356

Zheng R, Zhang C, Zhang S, Xu B (2014) Variational bayes based i-vector for speaker diarization of telephone conversations. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014, pp 91–95, DOI 10.1109/ICASSP.2014.6853564

Zhuang X, Tsakalidis S, Wu S, Natarajan P, Prasad R, Natarajan P (2012) Compact audio representation for event detection in consumer media. In: INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012, pp 2089–2092