# AN I-VECTOR PLDA BASED GENDER IDENTIFICATION APPROACH FOR SEVERELY DISTORTED AND MULTILINGUAL DARPA RATS DATA

*Shivesh Ranjan, Gang Liu, John H. L. Hansen*

Center for Robust Speech Systems (CRSS)
The University of Texas at Dallas, Richardson, TX, USA
$\{Shivesh.Ranjan, Gang.Liu, John.Hansen\}$@utdallas.edu

## ABSTRACT

This study proposes an i-Vector based approach to gender identification. Gender-labeled utterances from the Fisher English (FE) corpus are used to formulate an i-Vector extraction framework, and a Probabilistic Linear Discriminant Analysis (PLDA) back-end is employed to compute the scores for gender identification. A novel duration mismatch compensation strategy is also presented that offers very little degradation in identification accuracy even with a large reduction in the duration of the test-segment. The proposed method is shown to consistently outperform a GMM-UBM based gender-identification scheme on several test-sets created from a held-out portion of the FE corpus, and is able to achieve an identification accuracy of up to 97.63%. On the severely distorted and multilingual DARPA-RATS (Robust Automatic Transcription of Speech) corpora, the proposed approach achieves an identification accuracy of 76.48% using only the FE data in training. Next, a novel unsupervised domain adaptation strategy is also presented that utilizes only unlabeled RATS data to adapt the out-of-domain PLDA parameters derived from the FE training data. The strategy is able to offer a 6.8% relative improvement in identification accuracy, and a 14.75% relative reduction in Equal Error Rate (EER) compared to using the out-of-domain PLDA model on the RATS test-utterances. These improvements are significant since: 1) RATS test-utterances are severely distorted, 2) No labeled data of any kind is used for 4 of the 5 languages present in the test-utterances.

**Index Terms**: gender identification, i-vector, noisy, multilingual, duration mismatch, unsupervised domain adaptation

## 1. INTRODUCTION

Male and female speech differ due to a variety of factors, which can be primarily examined under three broad categories: physiological, acoustical and perceptual [1]. Several physiological parameters responsible for the differences between male and female speech have been identified, and examined. The primary reason for the differences between male and female speech are: difference in the vocal tract length, difference in length of vocal folds, and difference in the larynx anatomy [1].

Gender identification from speech is important for knowledge extraction to direct model selection/weighting in Automatic Speech Recognition (ASR), dialog systems, and diarization applications. While gender identification may be a somewhat solved problem for clean and monolingual corpora such as TIMIT, in extreme adverse conditions such as the DARPA RATS corpus, where the speech is both highly distorted and multilingual, gender identification is an extremely challenging problem. An automatic gender identification approach that can work on severely distorted and multilingual speech utterances, can serve a plenitude of purposes: improving speaker independent speech recognition accuracy; improving speaker-recognition accuracy. Gender identification algorithms are also used in accent identification, speaker health identification, emotion recognition, and in commercial applications such as: surveillance, call-center business applications, Human Computer Intelligent Interaction [1, 2, 3].

Several approaches for gender identification have been reported in the literature [1, 4, 5, 6]. In the gender identification approach outlined in [1], acoustic features including autocorrelation, linear prediction coefficients and others were used to form reference and test templates for vowels, fricatives, and unvoiced fricatives, and Euclidean distance was used in the identification experiments. In [4], a combination of scores using pitch estimation, and those from a Hidden Markov Model (HMM) system trained using Mel-scale based filter-bank coefficients, was used for gender identification. More recently, a Gaussian Mixture Model Universal Background Model (GMM-UBM) framework was used for gender identification in [5]. The use of GMM supervectors to train Support Vector Machines (SVM) for gender identification was also explored in the same work [5]. Very recently, a combination of several different strategies for gender and age identification like GMM-UBM posteriors based

This project was funded by AFRL under contract FA8750-12-1- 0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

scoring, GMM mean supervectors based SVM classification, sparse representation based on UBM weight posterior probability scores etc. were reported in [6].

i-Vector based systems are the current state-of-the-art in speaker verification, and offer a very effective way of representing speaker-specific models in a fixed dimensional space [7]. i-Vectors have also found widespread use in language identification [8]. Motivated by prior work on gender identification using the GMM-UBM framework, and the continued widespread use of i-Vectors in speaker verification/language identification tasks, this study presents a gender identification framework using i-Vectors.

The current study proposes an i-Vector based gender identification approach, trained using gender-labeled data from the Fisher English (FE) corpus, with a portion of the corpus randomly set aside for creating test-sets [9]. To identify the gender of a test utterance, the corresponding i-Vectors are extracted and scored against i-Vector models for both male and female speakers, using a Probabilistic Linear Discriminant Analysis (PLDA) back-end [10]. A novel duration mismatch compensation approach to handle shorter duration test segments is also presented. The proposed approach is shown to consistently outperform a GMM-UBM based gender identification approach on the FE test-sets of different duration.

The proposed gender identification approach is also considered on the severely noisy and degraded test utterances from the DARPA RATS corpora, with the test utterances from 8 different channels, and in 5 different languages (English, Pashto, Farsi, Urdu and Arabic). We also present a novel unsupervised domain adaptation strategy which uses unlabeled RATS data to adapt the out-of-domain PLDA back-end (derived from the FE training data) used in the gender identification.

This paper is organized as follows: Sec. 2 describes the proposed i-Vector PLDA based gender identification approach. A novel unsupervised domain adaptation strategy to adapt an out-of-domain PLDA model is presented in Sec. 3. Experiments, results and discussions are presented in Sec. 4, and the conclusions are in Sec. 5.

## 2. I-VECTOR BASED GENDER IDENTIFICATION

### 2.1. i-Vector Extraction

An i-Vector is a fixed low dimensional representation of a speech utterance that preserves the speaker-specific information. Since gender is important as a speaker-specific attribute, we hypothesize that gender information can be well represented by i-Vectors. In the i-Vector paradigm, a gender-specific GMM mean supervector $M$ can be represented in terms of the gender and channel independent supervector $m$, a low rank *total variability matrix* $T$, and a vector $w$ as

$$M = m + Tw. \tag{1}$$

In (1), $w$ is a random vector with a standard normal distribution $N(0, I)$. The $T$ matrix is learned using large amounts of labeled training data. In the gender-identification framework, utterance-labels are the gender of the corresponding speakers. The i-Vector of an utterance are its coordinates in the *total variability space* (i.e. space spanned by the columns of $T$), extracted as the maximum a posteriori (MAP) point estimates of $w$ given the utterance [11].

### 2.2. Probabilistic Linear Discriminant Analysis (PLDA) back-end for Gender Identification

After the training and test i-Vectors have been extracted, a variety of scoring techniques can be used to decide if they correspond to the same or different labels [7, 12, 13]. Here, we use a Probabilistic Linear Discriminant Analysis (PLDA) back-end for scoring which is the current state-of-the-art in speaker recognition systems.

Using i-Vectors extracted from a large labeled-training set, a PLDA model learns the within-class and across-class variabilities using an Expectation Maximization (EM) algorithm [10]. Specifically, we use a Gaussian PLDA (G-PLDA) of the form described in [12]. Assuming $R$ training utterances of a gender, an entire collection of gender-specific i-Vectors may be expressed as $\{\eta_r : r = 1, 2, ..., R\}$. In the G-PLDA parlance, an i-Vector of this collection can be expressed as,

$$\eta_r = m + \Phi\beta + \epsilon_r. \tag{2}$$

In (2), $m$ is a global offset, columns of $\Phi$ constitute a basis for the gender-specific subspace , $\beta$ corresponds to the coordinates in the gender-specific subspace, and $\epsilon_r$ is a Gaussian with zero mean and covariance $\Sigma$. The G-PLDA model parameters $\{m, \Phi, \Sigma\}$ are estimated using an EM algorithm on a large collection of gender-labeled training data. Given a test utterance, scores corresponding to competing hypotheses that belong to a female, or a male speaker are computed using the G-PLDA model, and a decision is made in favor of the hypothesis with the higher score. A closed form solution of G-PLDA model based score computation is given in [12], and is employed in the present work. For computing the scores, each gender is represented by a single i-Vector computed as the average of all training i-Vectors of the corresponding gender. A single gender-identification trial thus requires access to the two gender-specific i-Vectors, the test i-Vector, and the G-PLDA model parameters $\{m, \Phi, \Sigma\}$.

## 3. UNSUPERVISED DOMAIN ADAPTATION FOR OUT-OF-DOMAIN PLDA MODEL

The estimation of PLDA model parameters require large amounts of labeled training data for the corresponding domain. For many domains, such a large collection of labeled and balanced data may not be available, and in the extreme case, there may not be any labeled data for some domains. For example, one such specific case occurs when large amounts of gender-labeled data from Conversational Telephone Speech (CTS) is available to train an i-Vector PLDA based gender identification system, but no gender-labeled data from radio-channel speech is available. Such a system, trained entirely on CTS data may not perform well on test utterances from the radio-channel data. Recently, speaker recognition systems have focused on adapting out-of-domain PLDA models to a new domain with no labeled data [14, 15]. Such unsupervised domain adaptation strategies focus on using clustering algorithms to assign labels to unlabeled i-Vectors corresponding to a new domain, and then using these estimated labels to adapt the out-of-domain PLDA model to the new domain. The following subsections outline the clustering and adaptation steps involved in adapting an out-of-domain PLDA model to a new domain.

### 3.1. Unsupervised Binary Clustering

Unsupervised clustering techniques to perform domain adaptation for speaker recognition system were investigated in [15], and the performance of the speaker recognition system was shown to be sensitive to the choice of clustering algorithm used. In the same work, Agglomerative Hierarchical Clustering (AHC) was reported to offer better performance compared to other techniques. For a gender ID system, however, the number of desired clusters (i.e. 2) is much less compared to the number of clusters used in a typical speaker recognition system (AHC was used to create 1000 clusters in [15]).

We investigated 3 unsupervised clustering techniques: AHC, K-means and Label Generating-Max Margin Clustering (LG-MMC), and observed the gender ID system to be not very sensitive to the choice of the clustering algorithm. Label Generating-Max Margin Clustering (LG-MMC) is an unsupervised algorithm that operates by maximizing the margin of a 2-class dataset by generating the most violated labels iteratively, which are then combined using a multiple kernel learning strategy [16]. We adopted LG-MMC to assign labels to the unsupervised data since it can simultaneously learn both the optimal gender-labels, and the optimal separating hyperplane between the utterances corresponding to the two genders [16].

### 3.2. Model Adaptation for Out-of-domain PLDA Model

Once labels of the new domain's i-Vectors are estimated, we employ the same out-of-domain PLDA model adaptation procedure as described in [14]. Specifically, let $\Gamma$ be the across-class covariance matrix ($\Gamma = \Phi\Phi^T$), and $\Lambda$ be the within-class covariance matrix (same as $\Sigma$ of the G-PLDA model). Let $(\Gamma_{out}, \Lambda_{out})$ denote the parameters estimated using a resource rich out-of-domain gender-labeled dataset. Using the labels estimated by LG-MMC on the new domain's i-Vectors, a new set of parameters $(\Gamma_{in}, \Lambda_{in})$ is estimated. Next, a set of adapted parameters $(\Gamma_{adapt}, \Lambda_{adapt})$ is evaluated using,

$$\Gamma_{adapt} = \alpha_1 \Gamma_{in} + (1 - \alpha_1)\Gamma_{out} \tag{3}$$

$$\Lambda_{adapt} = \alpha_2 \Lambda_{in} + (1 - \alpha_2)\Lambda_{out}. \tag{4}$$

The parameters $\alpha_1, \alpha_2 \in [0, 1]$ appearing in (3) and (4) control the contribution of the in-domain, albeit initially unlabeled data to adapt the out-of-domain G-PLDA model. Figure 1 shows the setps involved in adapting an out-of-domain PLDA model to a new domain. The adapted parameters $(\Gamma_{adapt}, \Lambda_{adapt})$ are then used to perform a G-PLDA based scoring for gender identification on i-Vectors of the test utterances from the new domain.

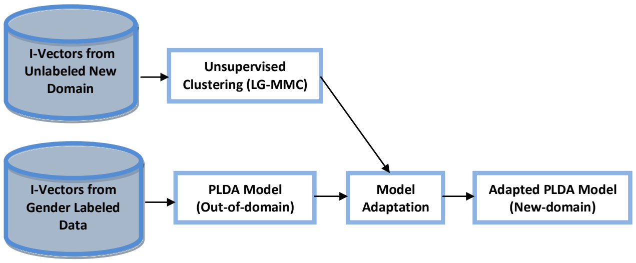
I-Vectors from Unlabeled New Domain
Unsupervised Clustering (LG-MMC)
I-Vectors from Gender Labeled Data
PLDA Model (Out-of-domain)
Model Adaptation
Adapted PLDA Model (New-domain)

**Fig. 1**. *Adapting the out-of-domain PLDA model to a new domain using unlabeled data.*

## 4. EXPERIMENTS, RESULTS AND DISCUSSIONS

### 4.1. Datasets

#### *4.1.1. Fisher English (FE) Datasets*

For training the proposed gender identification (ID) system, utterances from the FE corpus are used [9]. Specifically, approximately 11% of all FE utterances are randomly set aside for testing, and all remaining utterances used in training. A total of 20,652 utterances were used in training, and 2,600 utterances for testing. The average length per utterance after running an unsupervised Speech Activity Detection (SAD) algorithm was 234s [17]. Smaller training and test-sets were also created from the original sets described previously. We aimed to investigate the proposed gender ID system's performance across several test-sets with different duration. To this end, test-sets of duration 3s, 10s, and 20s were created from the original (complete) test-sets, by randomly selecting a single-segment of the desired duration per file.

i-Vector based speaker recognition systems are known to degrade when tested on shorter duration segments. To combat this issue, including shorter duration segments in training was suggested in [18]. We created separate training-sets of duration 3s, 10s, and 20s from the FE corpus by following the same procedure that was used for creating shorter duration test-sets. These training-sets were used later for testing shorter duration test-sets, using a novel duration mismatch compensation strategy (presented in Sec 4.4).

### *4.1.2. RATS Datasets*

The DARPA RATS program was aimed at developing several speech applications such as Keyword Spotting (KWS), SAD, Speaker Identification (SID), and Language Identification (LID) in severe noisy conditions. For the program, CTS data was retransmitted through eight different communication channels using multiple transmitters, receivers, and digitization equipment by the Linguistic Data Consortium (LDC). As a result of real-world radio transmission, the RATS data is severely degraded due to a variety of phenomena like high energy transmission bursts, nonlinear speech distortions, high channel noise, frequency shifts, and band limits [19].

For the RATS test-set used in this study, we used a total of 438 utterances from LDC2011R77, LDC2011E86, LDC2011E99 and LDC2011E111. A trained listener provided the gender-labels of the test files for the purpose of scoring. The average length per utterance was 816s. The test-set had data from all eight channels(A, B, C, D, E, F, G, H), and the noise-free source (SRC) utterances, with 49 utterances per channel for all channels and SRC, except H for which 46 utterances were present. Table 1 shows the composition of the RATS test in terms of the 5 languages present. Another dataset, with no gender-labels was also created using LDC2011R77, LDC2011E86, LDC2011E99 and LDC2011E111, for use in the out-of-domain PLDA model adaptation experiments. Specifically, 480 unlabeled utterances were used for channel H, and 502 utterances per channel for each of the remaining channels and the SRC.

| Language | No. of Test utterances |
|---|---|
| ENGLISH | 90 |
| PASHTO | 90 |
| FARSI | 78 |
| URDU | 90 |
| ARABIC | 90 |
| TOTAL | 438 |

**Table 1**. *Composition of the RATS test-set.*

## 4.2. System configuration

### *4.2.1. Proposed i-Vector PLDA based Gender ID System*

We use Mel-Frequency Cepstral Coefficients-Shifted Delta Coefficients (MFCC-SDC) as the acoustic features in all our experiments in this study [20]. MFCC-SDC are known to be very efficient in several acoustic event detections tasks such as LID and emotion recognition [20, 21]. It has been reported that male and female speech exhibit a gender-specific dynamic behavior due the difference in vocal tract dimensions [22]. We adopted the use of SDC features to utilize additional temporal information alongwith MFCC features, which may benefit from capturing the gender-specific dynamic behavior of speech. A common [7-1-3-7] configuration was used, with a window size of 25ms and a skip rate of 10ms to yield 56-dimensional features per frame [21]. The MFCC-SDC features were then used to train a UBM. For the UBM training, 4 iterations of the EM algorithm were used until the penultimate split, and 15 EM iterations were used in the final split. The same data used in UBM training was used to train the $T$ matrix using 5 iterations of the EM algorithm [7, 23].

### *4.2.2. GMM-UBM Baseline System*

The performance of our proposed i-Vector PLDA based gender ID system was compared against a standard GMM-UBM based approach. A MAP adaptation algorithm was used to adapt the UBMs developed during i-Vector training to obtain gender-specific GMMs [24]. The GMMs were then used to calculate the likelihood of the test utterances to assign gender labels.

## 4.3. Gender Separability in the i-Vector Space

We also investigated how genders are distributed in the i-Vector space, and if i-Vectors belonging to male and female speakers are clearly separated. To this end, a 3-dimensional projection of the 400-dimensional i-Vectors was obtained using Maximization of Mutual Information (MMI) based projection [25]. This MMI based projection has been shown to be better for representing several classes of high dimensional data in low dimensions, versus traditional techniques like Principal Component Analysis (PCA) or LDA [25]. Figure 2 shows a scatter plot of the first 2 dimensions of the MMI based projection for the FE test-set (complete duration). Clearly, gender information is well represented in the i-Vector space, with female (black) and male (red) speakers well separated.

## 4.4. Compensating Duration Mismatch for Gender Identification

We investigated the degradation in performance of the proposed i-Vector PLDA based gender ID approach when the duration of the test-segment is drastically reduced. To this end, we tested the system trained on complete FE utterances on FE test-sets of shorter duration. Table 2 shows the classification accuracy and EER obtained on 4 different test-sets. As can be seen, the accuracy decreases sharply from 97.62% for complete-duration test utterances to 65.58% for test-segments of duration 3s. A very drastic increase in corresponding EER can also be observed. It may also be observed that the degradation in performance is highly nonlinear, with the most severe degradation occurring when the test-segment's duration is reduced to 3s.

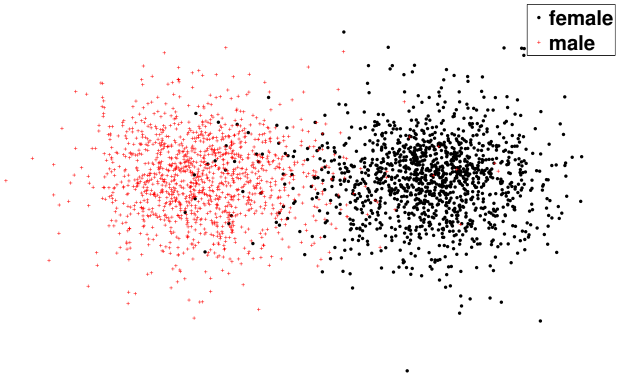

**Fig. 2**. *First 2 dimensions of MMI based projection of 400-dimensional i-Vectors of FE female (in black), and male (in red) test-set utterances.*

| Test Segment Duration | Accuracy (%) | EER (%) |
|---|---|---|
| Complete | 97.62 | 2.31 |
| 20s | 82.12 | 17.85 |
| 10s | 77.77 | 22.10 |
| 3s | 65.58 | 34.33 |

**Table 2**. *Degradation in performance of the proposed i-Vector PLDA based gender ID system due to duration mismatch. The gender ID system was trained on complete-duration FE training set.*

It was observed in [18] that including shorter duration segments in training can help to mitigate the effect of duration mismatch in a speaker recognition system. Along similar lines, we propose a novel strategy of training separate i-Vector PLDA systems such that the training and test duration segments are exactly matched. To keep the approach computationally efficient, we extracted only a single shorter duration training segment from the corresponding larger duration FE training utterance. While including more training data for shorter duration segments by creating multiple training segments per complete-utterance can certainly offer better gender ID performance, it will also lead to increased computational burden. Our simple, yet, novel strategy of using only a single short duration training segment per complete-utterance is highly effective, as we shall see subsequently (in Table 4).

Since we trained separate i-Vectors based systems corresponding to the different duration FE test-sets, both the number of components in the UBM, and the i-Vector dimensions were adjusted to account for the corresponding duration of the training utterances. Table 3 lists the configuration of duration mismatch compensated i-Vector PLDA based gender ID systems that were used on test-sets of corresponding duration.

| Train Duration | No. of mixtures in UBM | i-Vec Dim |
|---|---|---|
| 3s | 512 | 200 |
| 10s | 1024 | 200 |
| 20s | 1024 | 200 |
| Complete | 1024 | 400 |

**Table 3**. *Number of mixtures in the UBM, and i-Vector dimensions for the training-sets of different duration.*

### 4.5. Gender Identification Results

#### *4.5.1. Results on Fisher English Data*

Table 4 shows classification accuracy and EER obtained using our proposed approach compared against a GMM-UBM baseline system. For all the EERs reported in this study, the target class was female and Likelihood Ratio (LR) scoring was used. As can be seen, our approach consistently outperforms the GMM-UBM baseline in both metrics, and is able to achieve an accuracy of 97.62% on the complete duration test-set, and a low EER of 2.31%.

The difference in classification accuracy between our proposed system and the GMM-UBM baseline increases with increase in the duration of test (and training) set. It appears that with an increase in duration of test data, the i-Vectors are better able to utilize the additional information contained in the longer utterances than a GMM-UBM based framework.

It can also be observed that the identification accuracy and EER suffer very little degradation compared to the corresponding results in Table 2, demonstrating the efficacy of the novel duration mismatch compensation strategy presented earlier in sec 4.3. Since the test segments for the experiments reported in Table 2 and Table 4 are the same, a one-to-one comparison for the corresponding identification accuracy and EER of the i-Vector PLDA based gender ID system is valid. For the test-segments of durations 3s, the proposed domain mismatch compensation strategy offers an absolute gain of 25.69% in identification accuracy, and a 25.66% absolute reduction in EER compared to the uncompensated i-Vector PLDA based gender ID system.

| Test Duration | Accuracy (%) | | EER (%) | |
|---|---|---|---|---|
| | I-Vec PLDA | GMM UBM | I-Vec PLDA | GMM UBM |
| Complete | **97.62** | 95.23 | **2.31** | 4.46 |
| 20s | 96.15 | 94.62 | 3.85 | 4.69 |
| 10s | 94.85 | 93.65 | 5.15 | 6.23 |
| 3s | 91.27 | 90.73 | 8.67 | 9.00 |

**Table 4**. *Comparison of classification accuracy, and EER between the i-Vector PLDA based Gender ID approach and a GMM-UBM based system on test-sets from the FE corpus.*

### 4.5.2. Results on RATS Data

Our proposed gender ID system was also tested on utterances from the RATS test set. Since no gender-labeled RATS data is available, an unsupervised domain adaptation strategy to adapt the out-of-domain PLDA model (trained on the FE data) was also implemented to account for the domain mismatch.

Table 5 shows the channel-wise Gender ID classification accuracies, and EERs obtained using the out-of-domain FE-only trained i-Vector PLDA based system (shown as BEFORE), and that obtained after adapting the out-of-domain PLDA model using unlabeled RATS data (shown as AFTER). Clearly, the unsupervised domain adaptation approach gives significant improvement, as evident by an increase in the average classification accuracy by 5.25% (relative improvement of 6.8%), and a 3.08% (14.75% relative) reduction in the average EER.

| Channel | Accuracy (%) | | EER (%) | |
|---|---|---|---|---|
| | BEFORE | AFTER | BEFORE | AFTER |
| A | 61.22 | 79.59 | 32.65 | 20.41 |
| B | 73.47 | 79.59 | 23.47 | 20.41 |
| C | 77.55 | 77.55 | 21.43 | 21.43 |
| D | 75.51 | 77.55 | 19.39 | 21.43 |
| E | 61.22 | 61.22 | 37.76 | 37.76 |
| F | 81.63 | 89.80 | 17.35 | 10.20 |
| G | 95.92 | 100 | 4.08 | 0.00 |
| H | 65.22 | 73.91 | 29.35 | 26.09 |
| SRC | 95.92 | 95.92 | 3.06 | 3.06 |
| **AVG** | **76.48** | **81.73** | **20.89** | **17.80** |

**Table 5**. *Channel-wise classification accuracy, and EER of the i-Vector PLDA based Gender ID approach.* BEFORE *and* AFTER *refer respectively to results without, and with the adapted PLDA model (using unsupervised domain adaptation).*

Since the test-set has utterances from multiple channels with widely varying degradation and characteristics, we observe a large variation in performance for unsupervised model adaptation, when viewed channel-wise. Specifically, no change in classification accuracy is observed for channel C, E and SRC utterances, whereas large improvements are observed for channels A, F, and H. Here, the proposed model adaptation approach is not able to offer any improvements on the SRC test-utterances, since they correspond to clean CTS, which is similar (but not the same) to the FE corpora gender-labeled utterances. Thus, in this case, the out-of-domain PLDA model (estimated using FE data) is already very close to the adapted model. This hypothesis is also validated by the already high classification accuracy (95.92%), and low EER (3.06%) of the SRC test utterances. Moreover, the high SRC results also point to language-robustness of our proposed gender ID approach, as the SRC test utterances are from 5 different languages.

It appears that channel G test-utterances are gender identified perfectly (classification accuracy 100%) after model adaptation, as a result of some channel artifacts which when incorporated in the PLDA model (by the model adaptation) improve gender separability in the i-Vector space. Apparently, the nature of data from channels C and E prevent any improvements. The i-Vector system trained using the complete duration FE training-set was used for all the results in Table 5. We have used optimized model-adaptation parameters $\alpha_1, \alpha_2$ for the results, but it has been reported to be not a significant issue, as the model adaptation has been shown not to be sensitive around optimized values of the parameters as observed in [14, 15].

## 5. CONCLUSIONS

In this study, we presented an i-Vector PLDA based strategy for gender identification. We also presented a novel duration mismatch compensation approach that offered little degradation in gender identification accuracy and EER, even with a drastic reduction in the duration of the test-segments. Our proposed approach outperformed a GMM-UBM baseline on multiple test-sets created from the Fisher English corpus, and achieved classification accuracy and EER of up to 97.63% and 2.31% respectively.

We tested our approach on the severely distorted and multilingual DARPA RATS data, where no labeled data was available to adapt the out-of-domain PLDA model derived from a different corpus. We also presented a novel unsupervised domain adaptation strategy to adapt the out-of-domain PLDA model using only unlabeled data. Efficacy of this strategy is strongly validated by 6.8% relative gain in classification accuracy, and a 14.75% relative drop in EER compared to when only using the out-of-domain PLDA back-end. These improvements are significant since: 1) the RATS test data is highly degraded, with open-set languages, and from multiple low performance communication channels, while all the labeled training data is taken from the FE clean English corpus; 2) The RATS test data has utterances from 4 unseen languages (only English is common to our RATS test-set and the FE corpus). The proposed i-Vector PLDA based Gender Identification solution is highly effective for unseen noisy speech applications.

## 6. REFERENCES

[1] K. Wu and D. G. Childers, "Gender recognition from speech. part i: Coarse analysis," *Journal of Acoustical Society of America*, vol. 90, no. 4, pp. 1828–1840, 1991.

[2] I. Bisio, A. Delfino, F. Lavagetto, M. Marchese, and A. Sciarrone, "Gender-driven emotion recognition through speech signals for ambient intelligence appli-

cations," *IEEE Trans. Emerging Topics in Computing*, vol. 1, no. 2, pp. 244–257, 2013.

[3] F. Shahbaz, J. van de Weijer, M. A. Rao, M. Felsberg, and C. Gatta, "Semantic pyramids for gender and action recognition," *IEEE Trans. Image Processing*, vol. 23, no. 8, pp. 3633–3645, 2014.

[4] E. S. Parris and M. J. Carey, "Language independent gender identification," *IEEE ICASSP-1996*, vol. 2, pp. 685–688, 1996.

[5] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. Noth, "Age and gender recognition for telephone applications based on gmm supervectors and support vector machines," *IEEE ICASSP-2008*, pp. 1605–1608, 2008.

[6] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27, no. 1, pp. 151–167, 2013.

[7] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[8] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak, "Language recognition via i-Vectors and dimensionality reduction.," *ISCA INTERSPEECH*, pp. 857–860, 2011.

[9] C. Cieri, D. Graff, O. Kimball, D. Miller, and K. Walker, "Fisher English Training Parts 1 and 2, Speech and Transcripts," *Linguistic Data Consortium, Philadelphia*, 2005.

[10] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," *IEEE ICCV-2007*, pp. 1–8, 2007.

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

[12] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-Vector length normalization in speaker recognition systems.," *ISCA Interspeech*, pp. 249–252, 2011.

[13] G. Liu and J. H. L. Hansen, "An investigation into backend advancements for speaker recognition in multisession and noisy enrollment scenarios," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1978–1992, 2014.

[14] D. Garcia-Romero, A. McCree, S. Shum, N. Brummer, and C. Vaquero, "Unsupervised domain adaptation for i-Vector speaker recognition," *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[15] S. Shum, D. A. Reynolds, D. Garcia-Romero, and A. McCree, "Unsupervised clustering approaches for domain adaptation in speaker recognition systems," *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[16] Y.-F. Li, I. W. Tsang, J. T. Kwok, and Z.-H. Zhou, "Tighter and convex maximum margin clustering," *International Conference on Artificial Intelligence and Statistics*, pp. 344–351, 2009.

[17] L. N. Tan and A. Alwan, "Multi-band summary correlogram-based pitch detection for noisy speech," *Speech communication*, vol. 55, no. 7, pp. 841–856, 2013.

[18] T. Hasan, R. Saeidi, J. H. L. Hansen, and D. A. van Leeuwen, "Duration mismatch compensation for i-Vector based speaker recognition systems," *IEEE ICASSP-2013*, pp. 7663–7667, 2013.

[19] K. Walker and S. Strassel, "The RATS radio traffic collection system," *Odyssey*, 2012.

[20] M. A. Kohler and M. Kennedy, "Language identification using shifted delta cepstra," *Circuits and Systems, 2002. MWSCAS-2002. The 2002 45th Midwest Symposium on*, vol. 3, pp. III–69–72, 2002.

[21] G. Liu, Y. Lei, and J. H. L. Hansen, "A novel feature extraction strategy for multi-stream robust emotion identification.," *ISCA Interspeech*, pp. 482–485, 2010.

[22] A. P. Simpson, "Dynamic consequences of differences in male and female vocal tract dimensions," *Journal of Acoustical Society of America*, vol. 109, no. 5, pp. 2153–2164, 2001.

[23] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.

[24] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[25] K. Torkkola, "Feature extraction by non parametric mutual information maximization," *The Journal of Machine Learning Research*, vol. 3, pp. 1415–1438, 2003.