Integrating Online I-vector extractor with Information Bottleneck based Speaker Diarization system

Srikanth Madikeri, Ivan Himawan, Petr Motlicek and Marc Ferras

Idiap Research Institute , CH-1920 Martigny, Switzerland

srikanth.madikeri, ivan.himawan, petr.motlicek, mferras@idiap.ch

Abstract

Conventional approaches to speaker diarization use short-term features such as Mel Frequency Cepstral Co-efficients (MFCC). Features such as i-vectors have been used on longer segments (minimum 2.5 seconds of speech). Using i-vectors for speaker diarization has been shown to be beneficial as it models speaker information explicitly. In this paper, the i-vector modelling technique is adapted to be used as short term features for diarization by estimating i-vectors over a short window of MFCCs. The Information Bottleneck (IB) approach provides a convenient platform to integrate multiple features together for fast and accurate diarization of speech. Speaker models are estimated over a window of 10 frames of speech and used as features in the IB system. Experiments on the NIST RT datasets show absolute improvements of 3.9% in the best case when i-vectors are used as auxiliary features to MFCC. Further, discriminative training algorithms such as LDA and PLDA are applied on the i-vectors. A best case performance improvement of 5% in absolute terms is obtained on the RT datasets.

Index Terms: speaker diarization, online i-vectors, Information Bottleneck

1. Introduction

Speaker diarization addresses the problem of identifying *who spoke when* in a speech recording [1]. Speaker diarization systems based on the Hidden Markov Model/Gaussian Mixture Model (HMM/GMM) [2, 3] technique and the Information Bottleneck (IB) method [4] have been successfully applied on meeting data. On broadcast news recordings and telephone conversational recordings i-vector based approaches have been applied for the same task [5, 6]. The i-vector based approach has also been adapted to the meeting data in [7].

The IB approach to speaker diarization has been successfully employed in different conditions such as meetings, telephone conversations, etc. It provides a fast and convenient framework to combine multiple features by fusing posteriors from individual feature streams. Diarization systems typically operate at the short term feature level using features such as Mel Frequency Cepstral Co-efficients (MFCC). It is important that the features used and models obtained thereof represent speakers accurately.

In speaker recognition systems, the i-vector representation of a speaker has been shown to be successful [8]. An i-vector is a fixed-dimensional representation of a recording of speech from a single speaker, which can be further projected to a discriminative speaker space using techniques such as Linear Discriminant Analysis (LDA), Within Class Covariance Normalization (WCCN) and Probabilistic Linear Discriminant Analysis (PLDA) [9]. The success of i-vector systems has lead to its adoption to other problems where speaker analysis is required. In speaker diarization systems i-vectors have been used to analyse telephone conversation recordings. Subsequently, these methods have also been adapted to meeting conversations.

I-vectors are usually estimated on significantly long utterances. For instance, in speaker recognition datasets such as the NIST SRE datasets the average contribution of each speaker in a telephone conversation is 2.5 minutes. Diarization systems however perform analysis on short segments. Thus i-vector based approaches to speaker diarization have traditionally used a long context of speech. In one such approach ,the i-vectors are then clustered using the K-means algorithm [5]. Recent work from the authors have also demonstrated the use of SGMM (Subspace Gaussian Model) system based speaker vectors that are similar to the i-vectors in a similar framework [7]. Since the i-vectors are estimated over a long duration a realignment procedure is necessary to adjust the boundaries.

The diarization performance when using i-vectors suggest that they contain useful speaker information. Recent studies in Automatic Speech Recognition (ASR) have successfully explored adapting the i-vectors as short term feature representations. The i-vectors are appended to the short term MFCCs to train Deep Neural Network (DNN) based phone recognizers and decode the audio subsequently [10, 11]. In this paper, we study this method of a frame-level i-vector extraction when adapted as features to speaker diarization systems. As i-vectors are trained to represent speakers exclusively, we hypothesize that the feature representation will be useful for speaker diarization systems. It should be noted that in ASR the i-vectors are appended to MFCCs to train speaker-independent DNNs. In this work, the i-vectors are used in combination with MFCCs to discriminate speakers better. Discriminative algorithms such as LDA and PLDA are further applied to improve diarization performance.

The rest of the paper is organized as follows: Section 2 introduces the IB based diarization system briefly. In Section 4, the proposed system is described. The experimental results are presented in Section 5. Finally, the results are summarized in Section 6.

2. Information Bottleneck

The Information Bottleneck (IB) method diarizes an audio by optimizing the clusters with respect to a set of relevance variables [4]. The optimization criterion is given as follows:

$$\mathcal{F} = I(Y;C) - \frac{1}{\beta} I(C;X) \tag{1}$$

where X is the feature set, Y is the set of relevance variables and C is the set of clusters. β is the Lagrangian multiplier that controls the trade-off between information preserved in the

clusters and the cluster size. The term I refers to mutual information between two random variables.

The IB system is used to present our results throughout this work. The conventional IB system uses MFCC based features. Additionally, it may also use time domain information from the Time Delay Of Arrival (TDOA) features ([12]) for multiple distant microphone (MDM) recordings. Features such as Frequency Domain Linear Prediction (FDLP), Modulation Spectrum (MS) [13], Filterbank slope based features [14] have also shown to add complementary information. The IB framework for speaker diarization provides a fast and simple approach to combine multiple features without compromising on the performance. To combine multiple features the frame-level posteriors across the different feature streams are fused before IB clustering. Whereas in the HMM/GMM framework, which is a commonly used technique for speaker diarization, a common way to combine information from different features is to fuse the individual likelihood scores prior to Viterbi decoding [15]. This requires model re-estimation at every iteration and is thus computationally intensive.

In the IB framework, an audio recording is split into short segments of 2.5s. Each segment is parameterized by a multivariate Gaussian distribution estimated from the features with respect to the segment. The mean is computed from the segment and the covariance is computed from the entire utterance and shared across the mixtures. The mixtures are given weights based on the segment lengths. The parameters of the Gaussians are used to compute the posterior of each segment. The posteriors form the relevance variables Y in Equation 1. The agglomerative information bottleneck (aIB) clustering algorithm is applied to these segments [16, 17]. This is equivalent to the greedy optimization of Eq 1. The cost function simplifies to become the Jensen Shannon (JS) divergence between two clusters.

As the clustering is performed on fixed length segments, a final resegmentation step is applied using the Kullback-Leibler Hidden Markov Model (KL-HMM) segmentation algorithm. The posteriors for the KL-HMM algorithm are extracted with respect to the Gaussians estimated previously. For each segment, a mean posterior vector is extracted. The posterior of every frame is compared to these means using the KL divergence measure. Viterbi decoding on the sequence of posteriors is performed to get a new alignment of the speech frames. The overall KL divergence is minimized for Viterbi decoding. A minimum segment length constraint of 250 frames (for a frame rate at 100 frames per second) is applied while realigning.

3. Online i-vector extraction

The online i-vector extraction algorithm, as shown in Figure 1, extracts frame-level i-vectors from a stream of MFCC feature vectors. The i-vector extractor represents supervectors in a low-dimensional subspace. State-of-the-art speaker recognition systems use i-vectors to model speakers. For ASR, the i-vector technique have been recently employed for speaker adaptation using deep neural network. The speaker-specific information can be learned by stacking the i-vectors with acoustic features. This approach is shown to provide additional gains when used as an input to the DNN [18]. Another recent approach for speaker adaptation of DNN incorporates speaker i-vectors to project the speech features into a speaker-normalized space [19].

The i-vector framework follows the Total Variability Space (TVS) model, which is given by

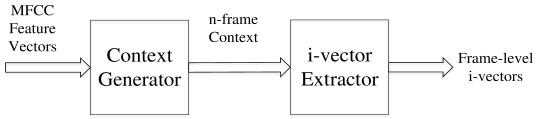

Figure 1: Block diagram representing the online i-vector extraction algorithm

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{2}$$

where $\mathbf{s}$ is the supervector adapted with respect to a Universal Background Model (UBM) from an utterance. The vector $\mathbf{m}$ is the mean of the supervectors, $\mathbf{T}$ is the matrix representing the subspace and $\mathbf{w}$ is the low-dimensional i-vector representation.

Given a sequence of MFCC feature vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t\}$, the first order statistics ($\mathbf{f}$) are estimated to get the i-vector representation. The subvector $\mathbf{f}_c$ of $\mathbf{f}$ is given by

$$\mathbf{f}_c = \mathbf{\Sigma}_c^{-\frac{1}{2}} \left(\sum \gamma_{t,c} \mathbf{x}_t - \boldsymbol{\mu}_c \right) \tag{3}$$

where $\mathbf{f} = [\mathbf{f}_1^t \mathbf{f}_2^t \ldots \mathbf{f}_C^t]^t$, C is the number of mixtures in the UBM, $\boldsymbol{\mu}_c$ and $\mathbf{\Sigma}_c$ are the mean and covariance matrix of the c^{th} mixture of the UBM. The posterior $\gamma_{t,c}$ for the t^{th} frame of speech with respect to the c^{th} mixture is given by

$$\gamma_{t,c} = P\left(\mathbf{x}_t | \boldsymbol{\mu}_c, \mathbf{\Sigma}_x\right) \tag{4}$$

The speech vectors are assumed to follow a Normal distribution.

Given the first order statistics, the i-vector is estimated as follows

$$\mathbf{w} = \left(\mathbf{I} + \sum_{c=1}^{C} N_c \mathbf{T}_c^t \Sigma_c^{-\frac{1}{2}} \mathbf{T}_c \right)^{-1} \mathbf{T}^t \mathbf{\Sigma}^{-1} \mathbf{f} \tag{5}$$

where $\mathbf{T}_c$ is the submatrix of $\mathbf{T}$ for the c^{th} mixture, $\mathbf{\Sigma}$ is the block diagonal matrix with $\mathbf{\Sigma}_c$ as blocks along the diagonal and

$$N_c = \sum_t \gamma_{t,c}$$

It is assumed that $\mathbf{T}$ has been whitened appropriately.

The online i-vector extraction algorithm uses the i-vector extraction procedure on a short window of MFCC features. As shown in Fig 1, a window of 10 frames is used (i.e. $t = 10$). For every set of 10 MFCC feature vectors, an i-vector is produced. This algorithm is called an online i-vector extractiona algorithm as the full audio is not required to extract an i-vector.

I-vectors obtained from a very short window of MFCC frames produce a feature stream similar to that of MFCCs and thus can be used as any other short term representation. It should be noted that the i-vector extractor is trained on similar short segments.

As i-vectors have been consistently shown to be good representations of speakers, using short term i-vectors can be useful in tasks that require speaker discrimination while operating on short segments. Speaker diarization system, thus, can benefit from these representations.

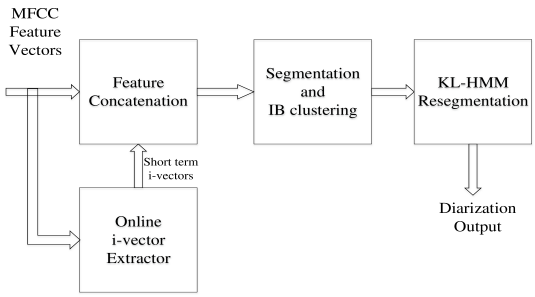

Figure 2: Block diagram representing proposed system

4. Proposed system

In this paper, the short term i-vector features are proposed to be used as a feature representation in the IB diarization system. The IB diarization system, as explained earlier, splits the audio into short segments (typically 2.5s). Each segment is a sequence of i-vectors. As the number of i-vectors is less than the number of short term feature (such as MFCC) due to the windowing of speech frames to extract i-vectors, the i-vectors are upsampled to match the speech features. The upsampling procedure simply involves replication of the i-vectors.

The upsample i-vectors are now used as a feature representation and each segment is represented by a Gaussian. Posteriors are extracted with respect to the estimated Gaussians for every i-vector. The mean of posteriors for each segment act as the reference variables for IB clustering. Next, the IB clusters are used to initialize the KL-HMM realignment algorithm.

4.1. Discriminative projection

Often in speaker recognition systems the i-vectors are passed through discriminative training algorithms as the i-vectors are only representation of the utterances and not speakers. Thus, discriminative training algorithms such as LDA and PLDA are used to further improve the representativeness of the speakers in the i-vector space. Multiple examples of short term i-vectors per speaker are used to train the LDA and PLDA parameters. The i-vectors used in the diarization system are then projected with these parameters. The projected i-vectors are used in the diarization system.

If $\mathbf{W}$ is the LDA projection matrix, $||.||$ represent the Euclidean norm of a vector, $\mathbf{F}$ and $\mathbf{E}$ are the PLDA parameters corresponding to the interclass and intraclass variances, the projection ($\hat{\mathbf{w}}$) is obtained as follows

$$\hat{\mathbf{w}} = \left(\mathbf{I} + \mathbf{F}^t\mathbf{E}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^t\mathbf{E}^{-1}\frac{\mathbf{W}\mathbf{w}}{||\mathbf{W}\mathbf{w}||} \quad (6)$$

As multiple examples of i-vectors are required to train the LDA and PLDA parameters, the training data is reused. The short-term i-vectors are used here as opposed to using i-vectors from the entire speech as often done in speaker recognition systems.

4.2. Feature fusion

The IB system provides a simple framework to fuse multiple features. Feature fusion involves estimating posteriors for each feature stream individually and fusing the resulting posteriors. It has been observed that fusing features is much more beneficial in the IB framework than in the HMM/GMM framework. In this paper, the fusion of i-vectors to the conventional MFCC feature streams is proposed. Even though the i-vectors are estimated from MFCCs, the two features provide different representations on the same data. While, the i-vector is estimated over a short window, the MFCCs represent single speech frames that are even shorter. Thus, the features are expected to be complementary. The architecture of the fused system is shown in Figure 2.

5. Experiments

Speaker diarization experiments are performed on the NIST RT 05, 06, 07 and 09 benchmark datasets. The NIST RT05 is used as a development dataset to tune PLDA parameters and fusion weights. The parameters are tuned to obtain the best Diarization Error Rate (DER). Multiple Distant Microphone (MDM) recordings are used for the experiments after their enhancement using *Beamformit* [20]. The proposed diarization system is compared with the IB based diarization system that uses only MFCC feature vectors. The i-vector PLDA system is the state-of-the-art technique in speaker recognition to model speakers. The open source Kaldi toolkit is used to train the online i-vector system and the parameters of LDA and PLDA. Speaker diarization is done using the IB diarization toolkit [21].

To compare the performance of the proposed system with the systems that use i-vectors in a longer context (eg: 2.5s), we include the results from [7] in which i-vectors are shown to perform better than the IB system. The i-vectors are clustered using K-means for every 2.5s of speech. The i-vector extractor is trained on the AMI dataset with MFCC features. The clustered segments are used as an initialization step to the KL-HMM segmentation algorithm. The posteriors for KL-HMM segmentation are obtained from the Gaussians estimated for every 2.5s long segments of the audio.

5.1. Feature extraction

MFCC features are extracted from the audio at 10ms frame rate with a window size of 25ms. A Gaussian is modelled for every 250 frames. The covariance matrix is shared across the Gaussians. The posteriors are estimated for every frame with respect to all the Gaussians.

5.2. Online i-vector system parameters

The open-source Kaldi speech recognition toolkit was used to train the online i-vector system [22]. The AMI and ICSI meeting data was used for training the TVS matrix. This dataset contains approximately 140 hours of segmented speech. During training, clean speech recordings captured by individual head microphone (IHM) with reference segmentation are used [23].

For the baseline system, the GMM/HMM system was trained on 39-dimensional MFCC features including delta and acceleration parameters. The acoustic models for the HMM/GMM systems have roughly 2.5K tied-states and a total of 100K Gaussians. The state alignments to train the DNN was obtained from the HMM/GMM system. The i-vector extractor used 13-dimensional MFCCs as the features with the frame length of 25 ms and shift of 10 ms. A 100-dimensional i-vector was generated for each utterance.

The i-vectors are upsampled to match the number of MFCC feature vectors for the speech segments in the audio. The silence segments are ignored for diarization.

Table 1: Results of experiments conducted on the NIST RT 05, 06, 07 and 09 datasets comparing presented. SER: Speaker Error Rate

	Dev. set	Test set		
System/Dataset	RT05	RT06 (SER)	RT07 (SER)	RT09 (SER)
MFCC	18.7	18.5	13.6	22.9
MFCC+ivec	**16.1**	20.7	**9.7**	**21.2**
MFCC+ivecPLDA	**16.5**	20.4	**8.6**	**21.3**

5.3. LDA and PLDA parameters

Speaker recognition systems apply discriminative algorithms on the i-vector to obtain better speaker representations. Usually, the i-vectors are projected on to a more discriminative space using LDA and PLDA. The LDA and PLDA models are trained on the short term i-vectors obtained from the training data used to train the i-vector extractor. We do not reduce the i-vector dimension after LDA or PLDA. It was observed (through the system performance on NIST RT 05 dataset) that reducing the dimensions hurt performance.

5.4. Results

The results on the NIST RT datasets are presented in Table 1. Three systems are compared: the baseline IB system using MFCC features only, IB system that uses MFCC and i-vectors and IB system that uses MFCC and i-vectors after PLDA projection as given in Equation 6. The results on the system using only i-vectors are not presented in the table as they are extremely poor. A diarization error rate (DER) of approximately 75% was obtained. The primary cause for this performance is the low frame rate for i-vectors (10 vectors per second as compared to 100 vectors in case of features such as MFCC). Even though the i-vectors are upsampled there is no new information gained.

The baseline system is obtained with MFCC features only. Based on the performance on NIST RT05 dataest it is clear that the i-vectors provide complementary information. The fusion weights are tuned on this dataset. Weights of 0.9 for MFCC and 0.1 for i-vector (or i-vector PLDA as the case may be) are observed to be optimal based on the diarization performance.

For the system using i-vector and MFCCs, a best case improvement of **3.9**% in absolute terms is obtained on the RT07 dataset. Improvements are obtained on all datasets except RT06. The application of LDA and PLDA to i-vectors improves the performance only for RT06 and RT07. The performance gains obtained are only significant on the RT07 dataset. The minimal gains obtained with PLDA as compared to that obtained in speaker recognition systems can be attributed to the severe mismatch in data across meetings in the RT datasets.

In Table 2, the results of the proposed systems are compared to the system that uses i-vectors estimated over longer segments. In particular, the system presented in [7] using i-vectors in the IB framework is used for comparison. The system uses RT07 and RT05 for development of parameters and hence are removed from comparisons. The comparison clearly shows that using online i-vectors is useful as compared to using i-vectors conventionally. On the RT09 dataset, an absolute improvement of 1.7% is observed on the MFCC+i-vec system. Using MFCC+i-vec-PLDA also provides improvements on the RT06 dataset, but the gains obtained are not significant enough to warrant the use of PLDA in diarization system, which are often required in real-time processing of speech signals.

Table 2: Results of experiments conducted on the NIST RT 06 and 09 datasets comparing the IB clustering and speaker vector (before and after PLDA) clustering methods SER: Speaker Error Rate, +PLDA: vectors projected in the PLDA space.

System/Dataset	RT06 (SER)	RT09 (SER)
Baseline (IB)	18.5	22.9
i-vector + PLDA ([7])	25.9	**21.3**
MFCC + Online i-vector	20.7	**21.2**
MFCC + Online i-vector + PLDA	20.4	**21.3**

6. Conclusion

Speaker diarization using short term features such as MFCC can further benefit from the short term i-vectors estimated using an online i-vector extraction algorithm. A best case improvement of 3.9% on the NIST RT 07 dataset corroborates our hypothesis. When compared to using i-vectors estimated over long speech segments in diarization systems, short term i-vector representations are observed to be more beneficial in the IB bottleneck framework.

7. Acknowledgements

This work was supported by the EU FP7 project Speaker Identification Integrated Project (SIIP) and the EC Eurostars project DBox: A generic dialog box for multi-lingual conversational applications.

8. References

[1] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Tran. on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[2] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2003*. IEEE, 2003, pp. 411–416.

[3] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Multimodal Technologies for Perception of Humans*. Springer, 2008, pp. 509–519.

[4] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," *IEEE Tran. on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1382–1393, 2009.

[5] S. Shum, N. Dehak, E. Chuangsuwanich, D. A. Reynolds, and J. R. Glass, "Exploiting intra-conversation variability for speaker diarization." in *INTERSPEECH*, 2011, pp. 945–948.

[6] S. Shum, N. Dehak, and J. Glass, "On the use of spectral and iterative methods for speaker diarization," in *Interspeech, Portland, Oregon*, 2012.

[7] S. Madikeri, P. Motlicek, and H. Bourlard, "Combining sgmm speaker vectors and kl-hmm approach for speaker diarization," in *To Appear In Proc. of ICASSP 2015*, 2015.

[8] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Tran. on Audio, Speech and Language Processing*, pp. 788–798, 2011.

[9] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE 11th International Conference on Computer Vision (ICCV)*. IEEE, 2007, pp. 1–8.

[10] H. N. D. P. M. Saon, G.; Soltau, "Speaker adaptation of neural network acoustic models using i-vectors," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, December 2013, pp. 55–59.

[11] V. Gupta *et al.*, "I-vector-based speaker adaptation of deep neural networks for french broadcast audio transcription," in *In Proc. of ICASSP 2014*, 2014.

[12] J. M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple distant microphone meetings: mixing acoustic features and inter-channel time differences." in *In Proc. of INTERSPEECH*, 2006.

[13] D. Vijayasenan, "An information theoretic approach to speaker diarization of meeting recordings," Ph.D. dissertation, ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE, 2010.

[14] S. Madikeri and H. Bourlard, "Filterbank slope based features for speaker diarization," in *In Proc. of ICASSP 2014*, 2014, pp. 111–115.

[15] X. Anguera, C. Woofers, and J. Hernando, "Speaker diarization for multi-party meetings using acoustic fusion," in *IEEE Workshop onAutomatic Speech Recognition and Understanding, 2005*. IEEE, 2005, pp. 426–431.

[16] N. Slonim and N. Tishby, "The power of word clusters for text classification," in *23rd European Colloquium on Information Retrieval Research*, vol. 1, 2001.

[17] D. Vijayasenan, F. Valente, and H. Bourlard, "Agglomerative information bottleneck for speaker diarization of meetings data," in *IEEE Workshop on Automatic Speech Recognition & Understanding, 2007*. IEEE, 2007, pp. 250–255.

[18] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 55–59.

[19] Y. Miao, H. Zhang, and F. Metze, "Towards speaker adaptive training of deep neural network acoustic models," in *Proc. Interspeech*, 2014.

[20] X. Anguera, "Beamformit (the fast and robust acoustic beamformer)." [Online]. Available: http://www.xavieranguera.com/beamformit/

[21] D. Vijayasenan and F. Valente, "Diartk: An open source toolkit for research in multistream speaker diarization and its application to meetings recordings." in *INTERSPEECH*, 2012.

[22] A. G. D. Povey *et al.*, "The kaldi speech recognition toolkit," in *In Proc. of ASRU 2011*, December 2011.

[23] T. Hain, L. Burget *et al.*, "Transcribing meetings with the amida systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 486–498, 2012.