VARIATIONAL BAYES BASED I-VECTOR FOR SPEAKER DIARIZATION OF TELEPHONE CONVERSATIONS

Rong Zheng, Ce Zhang, Shanshan Zhang, Bo Xu

Interactive Digital Media Technology Research Center
Institute of Automation, Chinese Academy of Sciences, Beijing, China
{rong.zheng, ce.zhang, shanshan.zhang, xubo}@ia.ac.cn

ABSTRACT

In this paper, we investigate the variational Bayes based I-vector method for speaker diarization of telephone conversations. The motivation of the proposed algorithm is to utilize variational Bayesian framework and exploit potential channel effect of total variability modeling for diarization of conversation side. Other three well-known techniques are compared as follows: K-means clustering for eigenvoices and I-vector speaker diarization, and variational Bayes applied to eigenvoices. Performance evaluations are conducted on the summed-channel telephone data from the 2008 NIST speaker recognition evaluation. The paper discusses how the performance is influenced by different modules, e.g., VAD, initial speaker clustering and Viterbi re-segmentation. Comparison experiments show the interest of variational Bayesian probabilistic framework for speaker diarization.

Index Terms—speaker diarization, eigenvoices, I-vector, total variability, variational Bayes

1. INTRODUCTION

Speaker diarization is the task of determining *who spoke when* in an audio recording, that is, to detect homogeneous audio segments with only one speaker and then group these segments with identical speaker label/cluster [1]. This paper addresses speaker diarization of telephone conversations where the number of speakers is known in advance.

As for speaker diarization, many research works are based on agglomerative and divisive hierachical manner such as top-down or bottom-up algorithms [2]. The bottom-up approach is by far the most popular system, that is, hierachical agglomerative clustering (HAC). Iterative process is used to merge similar segments. The authors in [3] compared the influence of five kinds of initialization algorithms for iterative-based speaker diarization. However, these greedy methods suffer from some limitations [4]. For example, the next merge is strongly dependent on previous clusters and error propagation at each step will increase the error rate.

In recent years, many different diarization algorithms have been proposed. Speaker diarization using factor analysis was first presented with an online stream-based method [5]. P.Kenny et al. presented variational Bayes based factor analysis (VB-based FA) and then compared with two systems, that is, agglomerative and soft speaker clustering, factor analysis-based diarization system [6]. The authors in [7] proposed I-vector based diarization approach examining intra-conversation variability in telephone speech. In [8], a global optimization framework based on I-vector paradigm was presented. An integer linear programming (ILP) solver was applied to minimize the result of diarization object function.

As shown in [6], eigenchannels were found to be ineffective in diarization of telephone conversations. However, special attention should be paid to this behavior. As we know, channel effect is usually helpful because it seems to distinguish two conversation sides, namely specific speaker and channel effect are contained in each conversation side. Another two facts also demonstrate this hypothesis. Firstly, raw (un-normalized) cepstral features provided better diarization performance comparing with popular normalized feature sets used in speaker recognition [6]. Secondly, I-vector based speaker diarization of telephone speech reported the same state-of-the-art performance as given by VB-based FA approach [7]. Here the total variability space potentially contains the speaker and channel variabilities simultaneously.

In this paper, we propose an approach to perform speaker diarization using variational Bayes and I-vector. Performance evaluations comparing our VB-based I-vector system with other well-known techniques are also provided.

The remainder of this paper is organized as follows. Section 2 describes a brief introduction to the total variability modeling. In section 3, we first introduce the variational Bayes for I-vector and then explain how to implement VB-based I-vector speaker diarization system. The experimental results are shown in section 4. Finally, some conclusions and discussions of relation to prior work are given in section 5 and section 6, respectively.

2. TOTAL VARIABILITY MODELING

Total variability modeling (also referred as I-vector) outlined by Dehak et al. [9] simultaneously models the speaker and channel variability in only one space. Let F and C be the acoustic feature dimension and the total number of Gaussian mixture components, the GMM-supervector is a CF -dimensional vector which is formed by concatenating the means of each mixture component. Given an utterance, the speaker- and channel- dependent GMM-supervector M , is written as,

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

where m is speaker- and channel- independent supervector, which is derived from UBM model, to represent the center of the full parameter space. T is a rectangular matrix of low rank and w is a standard normally distributed random vector. T and w are referred as total variability space and total factor, respectively.

This work is supported by Beijing Natural Science Foundation (No.4132071) and 973 Program of China (No.2013CB329302).

3. VARIATIONAL BAYES BASED I-VECTOR

3.1. The variational Bayes approach

For speaker recognition, the authors in [10] examined various sources of uncertainties in joint factor analysis (JFA) models from a Bayesian perspective. P.Kenny et al. showed that variational Bayesian framework could be applied to speaker diarization task [11]. The variational Bayes system, which incorporated eigenvoice and/or eigenchannel priors on GMM models, was proved to be very successful on the NIST 2008 summed channel data [6].

The variational Bayes algorithm is fully probabilistic approach which incorporates EM-like convergence guarantees and soft speaker clustering during the diarization process. Assuming that only one speaker is included when the initial speech segmentation is conducted, speaker diarization is formulated to calculate posterior probabilities of speaker s (S=2 for diarizaiton of telephone conversation in this paper) is talking in this short segment. These posterior probabilities were referred as segment posteriors in [6]. The diarization result is finally obtained by hard decision according to the maximal value of segment posteriors.

3.2. Variational Bayes for I-vector

The meaning of variational Bayes for speaker diarization is that a unified probabilistic framework is introduced to the model parameters of speaker GMM using some informative prior distributions for speaker- and channel-dependent GMM supervector. As for I-vector, we associate each speaker with a vector of total factors, that is, the variational Bayesian manner inspires us to implement total variability priors on GMM.

As for variational Bayes approach, each of the two speakers in the given telephone conversation is represented by a hidden vector of total factors, i.e., a posterior distribution in the total variability space. In order to calculate the segment posteriors, we should also calculate two speaker posteriors. For each of the two speakers in telephone conversation, the corresponding speaker posterior is defined as multivariate Gaussian distribution on total factors which models the location in the total variability space. This posterior distribution can be regarded as a point estimate of the GMM-supervector, according to equation (1), and the covariance matrix as a measure of the uncertainty of this point estimate in the total variability space.

The training procedures are listed as follows and the details of variational Bayes for I-vector are given in Algorithm 1:

Step1: Updating segment posteriors q_{rs}

Given the UBM model m and a sequence of L feature vectors $Y_r = \{y_1^r, y_2^r, ..., y_L^r\}$ for each segment r, the centered Baum-Welch statistics, γ_{rc} and f_{rc}, are calculated for mixture component c:

$$\gamma_{rc} = \sum_{t=1}^{L} p(c \mid y_t^r), \quad f_{rc} = \sum_{t=1}^{L} p(c \mid y_t^r)\, y_t^r - \gamma_{rc} m_c \tag{2}$$

Then q_{rs} is computed as,

$$\begin{aligned} \log \tilde{q}_{rs} &= const - \frac{1}{2} w_s^T \left(\sum_c \gamma_{rc} T_c^T \Sigma_c^{-1} T_c \right) w_s \\ &- \frac{1}{2} tr\left(Cov(w_s, w_s) \sum_c \gamma_{rc} T_c^T \Sigma_c^{-1} T_c \right) + \sum_c f_{rc}^T \Sigma_c^{-1} T_c w_s \end{aligned} \tag{3}$$

$$q_{rs} = \tilde{q}_{rs} / \sum_{s'} \tilde{q}_{rs'} \tag{4}$$

where *const* is independent of speaker s, Σ_c is the covariance matrix of the c- th component of UBM.

Step2: Updating speaker posteriors

$$Cov(w_s, w_s) = \left(I + \sum_c \gamma_c(s) T_c^T \Sigma_c^{-1} T_c \right)^{-1} \tag{5}$$

$$\langle w_s \rangle = Cov(w_s, w_s) \sum_c T_c^T \Sigma_c^{-1} f_c(s) \tag{6}$$

where $\langle w_s \rangle$ and $Cov(w_s, w_s)$ are the variational posterior mean and covariance matrix of w_s. $\gamma_c(s)$ and $f_c(s)$ are the centered zero and first order speaker dependent Baum-Welch statistics for mixture component c, which are the weighted sum of the segment dependent Baum-Welch statistics.

$$\gamma_c(s) = \sum_r q_{rs} \gamma_{rc}, \quad f_c(s) = \sum_r q_{rs} f_{rc} \tag{7}$$

Step3: Check the lower bound of the log-likelihood

$$\begin{aligned} L &= \sum_r \sum_s q_{rs} \log \tilde{q}_{rs} - \sum_r \sum_s q_{rs} \log q_{rs} + \frac{1}{2} R_t S \\ &+ \frac{1}{2} \sum_s \log |\, Cov(w_s, w_s) \,| - \frac{1}{2} \sum_s tr\left(Cov(w_s, w_s) + w_s w_s^T \right) \end{aligned} \tag{8}$$

where R_t is the rank of matrix T and S is the number of speakers in the conversation.

The segment and speaker posteriors are updated alternately until convergence is arrived. Then speaker diarization results can be obtained according to all of segment posteriors. For each segment r, we assign it to the speaker label s which provides the maximal segment posterior.

Algorithm 1 Procedure for VB based I-vector

Require: Segments $Y_1, ..., Y_r, ..., Y_R$

1: Accumulate the centered Baum-Welch statistics, γ_{rc} and f_{rc}, for each segment r
2: Randomly initialize w_s and q_{rs}, compute $L^{(0)}$
3: **for** k =1 **to** *nb_iterations* **do**
4: Update w_s, $Cov(w_s, w_s)$
5: Update q_{rs}, compute $L^{(k)}$
6: **if** $|L^{(k)} - L^{(k-1)}| \le \varepsilon$ **then**
7: stop
8: **end if**
9: **end for**

4. EXPERIMENTS AND RESULTS

4.1. Corpus

Experiments are conducted on the summed channel telephone data from the NIST 2008 speaker recognition evaluation (SRE). The corpus consists of 2213 telephone-quality conversations. Only two speakers are involved in each conversation, which is approximately five minutes in length.

4.2. Performance measurement

The performance assessment of speaker diarizatoin system is evaluated using Diarization Error Rate (DER). The DER is defined as time-weighted sum of three types of errors, i.e., miss detections, false alarms and speaker confusion errors [12]. As in [6,7], the standard deviation of the DER is also provided for reference.

In evaluating DER, if we use the reference speech/non-speech boundaries according to ASR transcriptions of SRE 2008 provided by NIST, it is shown that the maximal error is higher than 1000%. The reason is that only the first 12 seconds are transcribed in several conversations. So we obtain a reference VAD by applying a phoneme recognizer, which is borrowed from Brno University of Technology (BUT) [13], to each separate channel of the telephone conversation. The same reference speech activity detector was also used in [7].

4.3. Configuration

The first 20 Mel frequency cepstrum coefficients are extracted using a 25-ms Hamming window and a 10-ms frame rate, where any type of normalization and derivatives are not used. We train one gender-independent UBM with 1024 Gaussian components on a randomly selected subset from Fisher English part, Switchboard II Phase 2, Switchboard Cellular Part 2, NIST SRE 2004/2005/2006 data. The UBM is used to collect zero and first order Baum-Welch statistics. Then the total variability matrix of rank 100/300 is trained on the same corpus as UBM including 27531 recordings from 8131 female speakers and 22788 recordings from 6226 male speakers.

4.4. Voice activity detection

For initial speech segmentation, an energy-based GMM plus Viterbi decoding is used to generate the speech and non-speech boundaries. Based on normalized energy values, we initialize a 3-component GMM obtained by maximum likelihood criterion to model the distribution of three acoustic classes: low energy, high energy and silence. In the Viterbi stage, the component dependent likelihood is calculated to represent the emission probability of each HMM state. The self-loop transition probabilities are set to 0.6 for all states and we repeat low and high energy state for 10 times to constrain the switch frequency between speech and silence. Figure 1 shows the Viterbi stage of VAD module, which is defined as *automatic-VAD* in this work.

In order to avoid the degradation caused by initial speech segmentation and focus on the speaker confusion error only, we also use the reference boundaries provided by *reference VAD* for the next experiments.

After VAD stage, we simply remove the silences and chop the speech into 0.5 second intervals. A vector of speaker factors based on eigenvoices or total factors (I-vector) based on total variability matrix is extracted for each speech segment.

4.5. Initial speaker clustering

To perform the initial clustering step with eigenvoices or I-vectors, we use the following two strategies. Firstly, we simply use the K-means (K=2 for telephone conversations) clustering based on length-normalized cosine distance of speaker factors or I-vectors. Compared with hierachical agglomerative clustering with only one iteration to make hard decisions, K-mean clustering can adjust

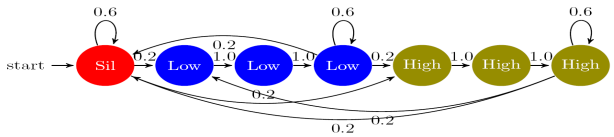

Figure1: Viterbi for VAD

deficient initialization with multiple iterations. However, K-means clustering actually makes hard decision within each iteration. So the second clustering strategy is the variational Bayes method described in section 3.2, which performs soft speaker clustering. In this work, Kmeans clustering and variational Bayes are regarded as different initial speaker clustering for comparisons.

4.6. Viterbi re-segmentation

Once the initial speaker clustering has been done, frame-based Viterbi re-segmentation is conducted to further improve the diarization result. We initialize three 32-component GMMs for three clusters, i.e., speaker A, speaker B and silence. Some experiments are examined to explore the influence of initial clustering after Viterbi re-segmentation.

4.7. Comparison systems

4.7.1. Eigenvoices (KM-EV)

Eigenvoices are used as prior knowledge of the speaker space, and then some low dimensional vector of speaker factors can be obtained for speaker diarization. K-means clustering is used as initial speaker clustering.

4.7.2. VB-based Eigenvoices (VB-EV)

The VB-based factor analysis diarization system was proposed in [6]. It was demonstrated that eigenchannels did not prove helpful in speaker diarization, so only eigenvoices are performed in this paper for comparison.

4.7.3. I-vector (KM-IV)

The i-vector system is based on the work described in [7]. After length normalization is applied, K-means speaker clustering is then conducted on the cosine distance.

4.7.4. The proposed VB based I-vector (VB-IV)

The proposed algorithm mentioned in section 3 is investigated. We report the experimental results and analyze the effects using our automatic VAD and reference VAD, respectively.

4.8. Results

First of all, we conduct experiments on eigenvoices and total variability matrix of rank 100. The performance difference with and without Viterbi re-segmentation are given, respectively. Then, the experiments are extended to matrix rank of 300. Finally, some validations are performed using reference-VAD.

4.8.1. Results on automatic-VAD

As shown in upper plot of Fig.2, without Viterbi re-segmentation, we remark that: (1) For K-means clustering, the I-vector diarization system is better than eigenvoices; (2) The diarization

performances of two variational Bayesian methods, i.e., VB-EV and VB-IV, are both substantially higher than those of K-means initial clustering, i.e., KM-EV and KM-IV; (3) However, when variational Bayes is used, the performance difference between eigenvoices and I-vector is neglectable.

After Viterbi re-segmentation, the performance gain of I-vector over eigenvoices is marginally small for K-means clustering and variational Bayes.

The diarization systems are then validated on eigenvoices and matrix T with rank 300 in lower plot of Fig.2. Comparing with rank 100, the diarization performances of rank 300 is more or less improved except that of VB with Viterbi. However, the general tendency is kept as found in rank 100. First of all, with or without Viterbi re-segmentation, KM-IV is better than KM-EV. Secondly, with variational Bayes is implemented, VB-EV is approaching to VB-IV, even though VB-IV is slightly better than VB-EV.

It seems that channel effect potentially contained by I-vector provides some perspective for speaker diarization, especially for K-means speaker clustering. Under the variational Bayesian framework, channel information shows marginal contribution to diarization, in contrast with slightly worse result reported by eigenvocies and eigenchannels against eigenvoices alone [6]. The comparable performances also show the interest of variational Bayesian probabilistic framework for speaker diarization.

4.8.2. Results on reference-VAD

We find that the initial speech segmentation by automatic VAD will highly influence the DER performance. For example, the performance improvement of speaker confusion error is smoothed out by increase of miss and false alarm errors. To avoid mismatched initial segmentation and compare our results with those systems described in [6,7], Table 1 and Table 2 provide experimental results using reference VAD boundaries, where only speaker confusion error is attributed to DER.

Because variational Bayes for rank 300 is time-consuming and performance of VB-IV for rank 100 is better, we only provide results of KM-EV and KM-IV on matrix rank 300 in Table 2. As given in Table 1, the best result given by variational Bayes based I-vector system obtains DER of 1.8% with a standard deviation of 5.31%, which is similar with that of variational Bayes based eigenvoices with a DER of 1.8% and standard deviation of 5.43%.

In view of the overall situation, firstly, the DERs have been obviously reduced comparing with the results shown in automatic VAD (refer to Fig.2). Secondly, the proposed system obviously benefits from the variational Bayesian framework. Thirdly, our experiments also provide different perspective on channel effect for diarization according to the implementation of initial speaker clustering strategies. For K-means speaker clustering, channel effect is proved to be helpful, while the effectiveness is evidently restrained under the variational Bayesian framework.

5. CONCLUSIONS

In this paper we examine variational Bayes based I-vector for speaker diarization over telephone conversations. We introduce the total variability prior distribution with Bayesian approach to perform speaker diarization. Iterative refinements of the approximate posterior probability related with hidden random variable are calculated, which is guaranteed to increase the lower bound on model log-likelihood. We compare it with other popular techniques, e.g., eigenvoices, I-vector, variational Bayes based eigenvoices. Comparison experiments provide some perspective on variational Bayesian framework for speaker diarization. Relative effectiveness of channel effect is also reported in our experiments.

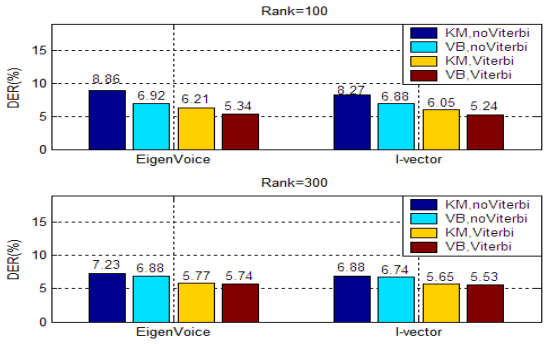

Figure 2: Comparisons of diarization results of comparative systems on the NIST-SRE 2008 summed channel telephone data. All systems are initialized using automatic VAD. The rank of eigenvoices and total variability matrix are set to 100 and 300, respectively. (KM: KMeans clustering; VB: variational Bayes).

Table 1: *Comparisons of diarization results using reference VAD. The rank of eigenvoices and total variability matrix are both set to 100. (Stdev: standard deviation of DERs; KM-EV: eigenvoices with Kmeans; KM-IV: I-vector with Kmeans; VB-EV: variational Bayes based EV; VB-IV: variational Bayes based IV).*

		DER(%)	Stdev(%)
KM-EV	No Viterbi	4.8	7.47
	Viterbi	3.0	6.67
VB-EV	No Viterbi	2.5	5.80
	Viterbi	**1.8**	**5.43**
KM-IV	No Viterbi	4.1	7.12
	Viterbi	2.7	6.55
VB-IV	No Viterbi	2.5	5.69
	Viterbi	**1.8**	**5.31**

Table 2: *Comparisons of diarization results using reference VAD. The rank of eigenvoices and total variability matrix are set to 300.*

		DER(%)	Stdev(%)
KM-EV	No Viterbi	4.1	5.81
	Viterbi	2.4	5.64
KM-IV	No Viterbi	2.7	6.31
	Viterbi	2.2	6.09

6. RELATION TO PRIOR WORK

The I-vector speaker diarization proposed in [7] was designed with hard decisions using K-means initial speaker clustering, we develop I-vector using the variational Bayes, which performs soft speaker clustering. For the diarization task where the number of participating speakers is unknown, the authors of [7] used the variational Bayesian GMM for model selection and applied it to estimate the number of speakers [14].

A similar work on variational Bayes for speaker diarization was described in [6], which incorporated eigenvoice priors for diarization. However, eigenchannels did not prove helpful in their experiments. Considering channel effects give more realistic and useful prior to distinguish two conversation sides, the present study is related to variational Bayes and focuses on the total variability space potentially contains speaker and channel variability.

7. REFERENCES

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

[2] N. Evans, S. Bozonnet, D. Wang, C. Fredouille, and R. Troncy, "A comparative study of bottom-up and top-down approaches to speaker diarization," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 2, pp. 382–392, 2012.

[3] O. Ben-Harush, I. Lapidot, and H. Guterman, "Initialization of iterative-based speaker diarization systems for telephone conversations," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 2, pp. 414–425, 2012.

[4] I. Lapidot and H. Guterman, "Future challenges in speaker diarization," Proc. of Speech Processing Conference, 2011.

[5] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices," in Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on. IEEE, 2008, pp. 4133–4136.

[6] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," Selected Topics in Signal Processing, IEEE Journal of, vol. 4, no. 6, pp. 1059–1070, 2010.

[7] S. Shum, N. Dehak, E. Chuangsuwanich, D. Reynolds, and J. Glass, "Exploiting intra-conversation variability for speaker diarization," in Proc. Interspeech, 2011.

[8] M. Rouvier and S. Meignier, "A global optimization framework for speaker diarization," in Odyssey 2012-The Speaker and Language Recognition Workshop, 2012.

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

[10] X. Zhao and Y. Dong, "Variational bayesian joint factor analysis models for speaker verification," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 3, pp. 1032–1042, 2012.

[11] P. Kenny, "Bayesian analysis of speaker diarization with eigenvoice priors," CRIM, Montreal, Technical Report, 2008.

[12] DER scoring code. Available: www.nist.gov/speech/tests/rt/2006-spring/code/md-eval-v21.pl

[13] Phoneme recognizer based on long temporal context. Available: http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context

[14] S.Shum, N.Dehak, R.Dehak, and J.Glass, "Unsupervised methods for speaker diarization: an integrated and iterative approach," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 21, no. 10, pp. 2015–2028, 2013.