VARIATIONAL BAYESIAN PLDA FOR SPEAKER DIARIZATION IN THE MGB CHALLENGE

Jesús Villalba, Alfonso Ortega, Antonio Miguel, Eduardo Lleida

ViVoLab, Aragon Institute for Engineering Research (I3A),
University of Zaragoza, Spain
{villalba,ortega,amiguel,lleida}@unizar.es

ABSTRACT

This paper describes the ViVoLab speaker diarization system for the Multi-Genre Broadcast (MGB) Challenge at ASRU2015. The challenge data consisted of BBC TV programmes of different genres. Diarization followed a longitudinal setup, i.e., the speakers of the current episode had to be linked to the speakers in previous episodes of the same show. We propose a system based on the i-vector paradigm. After an initial segmentation step, we compute an i-vector per speech segment. Then, a generative model based on Bayesian PLDA clusters the speakers. In this model, the speaker labels are latent variables that we optimize by variational Bayes iterations. The number of speakers in each episode was decided by maximizing the variational lower bound. The system includes several phases of segment-merging and re-clustering. We re-compute i-vectors after each merging step, which reduces the i-vector uncertainty. This approach attained a DER around 30% in the development set.

Index Terms— Speaker Diarization, i-vectors, PLDA, variational Bayes, MGB challenge

1. INTRODUCTION

The Multi-Genre Broadcast (MGB) Challenge at ASRU 2015 is an evaluation of speech recognition and diarization systems on BBC TV recordings [1]. It includes multiple TV shows of different genres. The evaluation consists of four tasks: speech-to-text transcription, subtitle alignment, longitudinal speech-to-text transcription and longitudinal speaker diarization. In this paper, we present the diarization system that we submitted to the diarization track. This task followed a longitudinal setup, i.e., to process one episode we can use information of previous episodes of the same show, and we have to link the speakers between episodes. Only the data provided by the organizers could be used to train background models, PLDA, etc. This fact is challenging since most of this data did not include reliable speaker labels.

This work has been supported by the Spanish Government through project TIN2014-54288-C4-2-R and by the European Unions's FP7 Marie Curie action, IAPP under grant agreement no. 610986.

Speaker diarization is a problem related to speaker recognition that determines "who spoke when". The interest for this topic has rapidly increased in the last years. It has application on telephone conversations, meetings, broadcast news, movies, etc. Besides, the explosion of multimedia content in the Internet requires of automatic means to index that information. Detailed reviews about the diarization technology evolution can be found in [2, 3].

The first successful diarization systems were bottom-up approaches based on segmentation of acoustic features by Bayesian Information Criterion (BIC) [4] followed by Agglomerative Hierarchical Clustering (AHC) [5]. With the advent of joint factor analysis (JFA) [6], we find approaches based on clustering streams of speaker factors [7, 8]. These approaches compute speaker factors using a sliding window of about 1 second length. The speaker factors are, then, clustered combining several metrics and algorithms like PCA, K-means, GMM, etc. Also based on JFA, we find approaches using variational Bayes (VB) [9, 10]. In these approaches, both speaker factors and speaker labels are latent variables that are jointly estimated by maximizing a lower bound on the data likelihood.

In this work, we mix stream methods and variational Bayes. First, we extract a stream of i-vectors [11] from the speech file. Then, variational Bayes PLDA clusters those i-vectors into speakers. This approach was used successfully for training PLDA models with unlabeled datasets [12, 13].

This paper is organized as follows. Section 2 describes the structure of our speaker diarization system. Sections 3 to 5 explain the building blocks of the system: initial segmentation, i-vector extraction, and clustering based on VB PLDA. Section 6 describes the experimental setup, including the dataset, and it details the configuration of the system. Section 7 discusses the results on the development and evaluation data. Finally, Section 8 summarizes the conclusions.

2. DIARIZATION SYSTEM DESCRIPTION

Figure 1 shows the diagram of our longitudinal speaker diarization system. The system receives the speech signal $s(n)$ of the current episode, the i-vectors $\mathbf{\Phi}_0$ corresponding to the

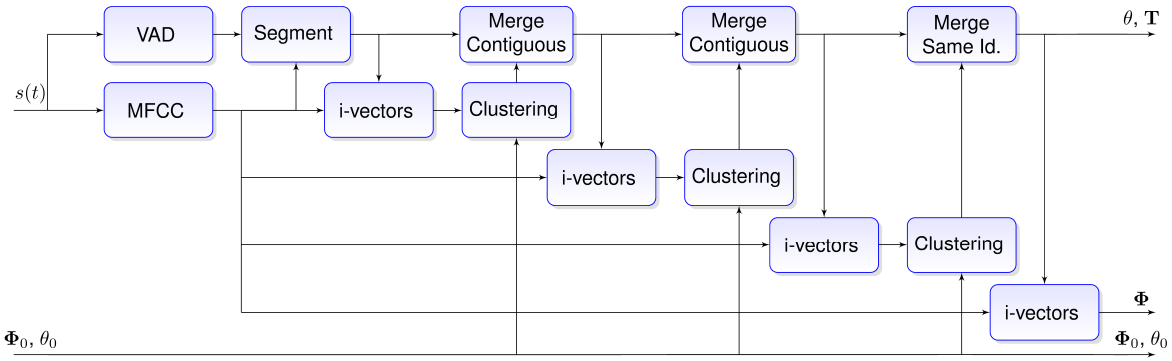

Fig. 1: Block diagram of the longitudinal speaker diarization system.

speakers in previous episodes of the show, and the speaker labels associated to them θ_0. The outputs of the system are the final segments with their time stamps $\mathbf{T}$, i-vectors for each of the speakers in the episode $\mathbf{\Phi}$ with their speaker labels θ, and the i-vectors and labels of the previous episodes.

The system uses MFCC as features. We used the baseline VAD provided by the organizers. We perform an initial segmentation based on Bayesian Information Criterion (BIC). The purpose of this step is to find changes of the speakers' turns in the episode. Thus, we obtain many short-segments containing a single speaker. Then, we compute an i-vector per speech segment. The i-vectors are clustered into speakers by variational Bayes PLDA, explained in detail in Section 5. We add the i-vectors of the previous episodes to carry out the speaker linking between episodes.

With the labels of the first clustering, we merge together contiguous segments that have been assigned to the same speaker. After that, we recompute the i-vectors. The reason to do this is that i-vectors computed with longer segments have less uncertainty and more discriminant power. Then, we cluster the new i-vectors again. We repeat the process of merging adjacent segments, recompute i-vectors and reclustering once more to obtain the final clustering. The final step consists in merging all the segments with the same speaker id. and compute one i-vector per speaker. We pass these i-vectors together with the i-vectors of previous episodes to the next episode.

3. BIC SEGMENTATION

The initial speaker change points are detected using a Bayesian Information Criterion (BIC) distance metric [4, 5]. This method searches for change points in a window of features. The window grows until a change point is detected. Then, the window is reset to start in the change point and the search for the next change point starts again. For each point in the window, we compute a penalized likelihood ratio test between the hypothesis (H_0) that the window is better modeled by two distributions, one for each side of the change point candidate; and the hypothesis (H_1) that the window is better modeled by a single distribution. That is the difference of BIC values,

$$\Delta\text{BIC} = \log(R) - \lambda P \tag{1}$$

where R is the likelihood ratio between both hypothesis, P is a penalty term that measures the excess of complexity of H_0 w.r.t. H_1 and λ is a scaling hyperparameter. Full covariance Gaussians were used to model the windows distributions.

4. I-VECTOR EXTRACTION

The i-vector paradigm [11] is an extension of the GMM-UBM approach [14], where a speech segment is modeled by a Gaussian mixture model (GMM). The i-vector approach assumes that the super-vector mean $\mathbf{M}$ of the segment GMM can be written as

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\phi \tag{2}$$

where $\mathbf{m}$ is the UBM means super-vector, $\mathbf{T}$ is a low-rank matrix and ϕ is a standard normal distributed vector. $\mathbf{T}$ defines the total variability space, i.e. the directions in which we can move the UBM to adapt it to a specific segment.

Using this model, we can compute the posterior distribution of ϕ given the segment data. This posterior is Gaussian distributed and the mean of this distribution is referred as the i-vector in the literature. Thus, we can model a sequence of variable length with a single feature vector. The i-vector becomes a new feature for pattern classification algorithms like SVM [15] and PLDA [16, 17].

It has been observed that normalizing the i-vector by its magnitude–known as length normalization–improves the discriminant power of the i-vector [18]. Before length normalization, we center and whiten the i-vectors to assure that they are evenly distributed in the unit hypersphere.

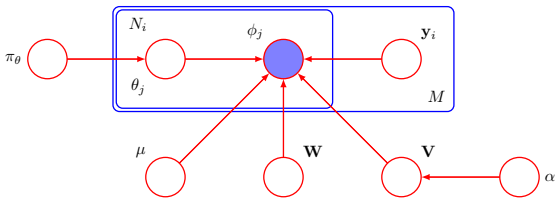

Fig. 2: BN for unsupervised SPLDA.

5. CLUSTERING WITH UNSUPERVISED PLDA

5.1. Model description

We introduced the Unsupervised PLDA model in [12, 13]. In those works, this model was used to train probabilistic linear discriminant analysis (PLDA) on a dataset with unknown speaker labels. Speaker clustering and model parameters estimation was performed iteratively. Here, we apply this model to the speaker clustering phase of our diarization system.

The model is based on Simplified PLDA (SPLDA), that is a linear generative model that assumes that an i-vector ϕ_j of speaker i can be written as:

$$\phi_j = \mu + \mathbf{V}\mathbf{y}_i + \epsilon_j \tag{3}$$

where μ is a speaker independent term, $\mathbf{V}$ is a low rank eigenvoices matrix, $\mathbf{y}_i$ is the speaker factor vector, and ϵ_j is the within class variability term. We put a standard normal prior on $\mathbf{y}_i$ and normal with zero mean and precision $\mathbf{W}$ on ϵ_j.

Figure 2 depicts the Bayesian network of this model where the speaker labels θ of the data are hidden. θ partitions N i-vectors into M speakers. θ_j is a latent variable comprising a 1–of–M binary vector with elements θ_{ji} with $i = 1, \ldots, M$. Note that the distribution of each speaker is assumed to be Gaussian with mean $\mu + \mathbf{V}\mathbf{y}_i$ and precision $\mathbf{W}$. The set of all the speakers forms a GMM where θ corresponds to the component occupations. The conditional distribution of θ given the mixture weights π_θ is

$$P(\theta|\pi_\theta) = \prod_{j=1}^{N}\prod_{i=1}^{M} \pi_{\theta_i}^{\theta_{ji}} \ . \tag{4}$$

We put a Dirichlet prior on the weights:

$$P(\pi_\theta|\tau_0) = \mathrm{Dir}(\pi_\theta|\tau_0) = C(\tau_0)\prod_{i=1}^{M} \pi_{\theta_i}^{\tau_0 - 1} \tag{5}$$

where, by symmetry, we choose the same τ_0 for all the components, $C(\tau_0)$ is the normalization constant,

$$C(\tau_0) = \frac{\Gamma(M\tau_0)}{\Gamma(\tau_0)^M} \tag{6}$$

and Γ is the Gamma function.

5.2. Model priors

In [12], we compared two versions of this model: one where the model parameters were point estimates and another where the model parameters are hidden variables with prior and posterior distributions. The latter–using weak informative priors–produced PLDA models with lower error rates. Here again, we use the fully Bayesian version where μ, $\mathbf{V}$ and $\mathbf{W}$ are latent variables.

We put a *hierarchical* prior $P(\mathbf{V}|\alpha)$ over $\mathbf{V}$ [19] through a conditional Gaussian distribution of the form:

$$P(\mathbf{V}|\alpha) = \prod_{q=1}^{n_y} \left(\frac{\alpha_q}{2\pi}\right)^{d/2} \exp\left(-\frac{1}{2}\alpha_q \mathbf{v}_q^T \mathbf{v}_q\right) \tag{7}$$

where $\mathbf{v}_q$ are the columns of $\mathbf{V}$ and n_y is the speaker factors dimension. Each α_q controls the inverse variance of the corresponding $\mathbf{v}_q$. If a particular α_q has a posterior distribution concentrated at large values, the corresponding $\mathbf{v}_q$ will tend to be small, and that direction of the latent space will be effectively "switched off".

We defined a prior for α:

$$P(\alpha) = \prod_{q=1}^{n_y} \mathcal{G}(\alpha_q|a_\alpha, b_\alpha) \tag{8}$$

where $\mathcal{G}$ denotes the Gamma distribution.

We placed a Gaussian prior for the mean μ:

$$P(\mu) = \mathcal{N}\left(\mu|\mu_0, \beta^{-1}\mathbf{I}\right) \ . \tag{9}$$

Low values of a_α, b_α and β make the priors less informative and vice versa.

Finally, we put a Wishart prior on $\mathbf{W}$,

$$P(\mathbf{W}) = \mathcal{W}(\mathbf{W}|\mathbf{\Psi}_0, \nu_0) \ . \tag{10}$$

5.3. Variational Bayes approximation

The equations of this model cannot be solved in close form so we used a variational Bayes approximation. We approximated the joint posterior of the latent variables by a factorized distribution of the form:

$$P(\mathbf{Y}, \theta, \pi_\theta, \mu, \mathbf{V}, \mathbf{W}, \alpha|\mathbf{\Phi}) \approx$$
$$q(\mathbf{Y})\, q(\theta)\, q(\pi_\theta) \prod_{r=1}^{d} q(\tilde{\mathbf{v}}_r')\, q(\mathbf{W})\, q(\alpha) \tag{11}$$

where $\tilde{\mathbf{v}}_r'$ is a column vector containing the r^{th} row of $\tilde{\mathbf{V}} = [\mathbf{V} \quad \mu]$. If $\mathbf{W}$ were diagonal the factorization $\prod_{r=1}^{d} q(\tilde{\mathbf{v}}_r')$ would not be necessary because it would arise naturally. However, for full $\mathbf{W}$, we have to force the factorization to make the problem tractable.

We computed these factors by using variational Bayes [20] with deterministic annealing (DA) [21]. The formula to update a factor q_l is

$$\ln q_l^* (\mathbf{Z}_l) = \mathrm{E}_{m \neq l} [\kappa \ln P (\mathbf{\Phi}, \mathbf{Z})] + \text{const} \qquad (12)$$

where $\mathbf{Z}$ abbreviates the set of all hidden variables, $\mathbf{Z}_l$ are the hidden variables corresponding to the l^{th} factor, and κ is the annealing factor; expectations are taken with respect to all the factors $m \neq l$. We can prove that Equation (12) optimizes the VB lower bound

$$\begin{aligned} \mathcal{L} &= \mathrm{E} [\ln P (\mathbf{\Phi}, \mathbf{Z})] - \mathrm{E} [\ln q (\mathbf{Z})] \\ &= \ln P (\mathbf{\Phi}) - \mathrm{KL} (q (\mathbf{Z}) || P (\mathbf{Z}|\mathbf{\Phi})) \end{aligned} \qquad (13)$$

where expectations are taken with respect to the variational posterior $q (\mathbf{Z})$. $\mathcal{L}$ is an approximation of the marginal likelihood of the data $\ln P (\mathbf{\Phi})$, which becomes equality when approximated posterior is equal to the true posterior. Annealing modifies the VB objective in a way that helps to avoid local maxima. We must set $\kappa < 1$ at the beginning and increase it as $\kappa \leftarrow 1.1\kappa$ in each iteration until $\kappa = 1$.

The full VB equations can be found in our report [22].

5.4. Initialization with AHC

The VB algorithm needs some initialization for the speaker labels. We initialized them with Agglomerative Hierarchical Clustering (AHC) [23]. AHC is a greedy bottom-up approach. Initially, each i-vector is its own cluster and, then, clusters are progressively merged using a similarity criterion. Thus, we start with the pair-wise score matrix between all the development i-vectors. This score matrix is obtained with an initial PLDA model or cosine similarity. Then, we use a linkage criterion to determine the similarity between the clusters A and B, $s(A, B)$, as a function of the pair-wise scores between their elements $s(a, b)$. In [13], we tried several linkage criteria (average, complete and single) obtaining better results with the average criterion,

$$s_{\text{avg}}(A, B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} s(a, b) \; . \qquad (14)$$

5.5. Model selection

This model requires to hypothesize, the number of speakers in the dataset. As the number of speakers M is unknown, we ran the AHC+VB algorithm several times, each time hypothesizing a different M. We assumed that the best model is the one that obtains the largest VB lower bound $\mathcal{L}(M)$. To fairly compare lower bounds for different M, the Dirichlet prior on the speaker weights needs to be such that the product $M\tau_0$ is constant. We do several iterations to find the best M, in each iteration we try 5 different M values. Depending on the resulting lower bounds, we choose the M values to try in the next iterations, until we find the optimum value for M.

5.6. Applications of the model

5.6.1. Unsupervised PLDA training

We can use this model to train PLDA on a dataset with unknown labels. In this work, we trained PLDA combining the development data, which was labeled, and the training data, which was unlabeled. The variational factors of the model parameters (μ, $\mathbf{V}$, $\mathbf{W}$ and α), the speaker factors of the training and development set ($q (\mathbf{Y}_{\text{DEV}}) q (\mathbf{Y}_{\text{TRN}})$) and the labels and label priors of the training set ($q (\theta_{\text{TRN}})$ and $q (\pi_{\theta_{\text{TNR}}})$) are updated iteratively. Besides, we did not allowed common speakers between different shows, but we allowed common speakers between episodes of the same show.

5.6.2. Diarization clustering

In the clustering steps of the diarization system, we only update the variational posteriors corresponding to the speaker factors $q (\mathbf{Y})$, and the speaker labels $q (\theta)$ and the label priors $q (\pi_\theta)$. The variational factors corresponding to the PLDA model parameters were not updated, we kept the ones obtained in the training step.

6. EXPERIMENTAL SETUP

6.1. MGB Challenge dataset

The MGB Challenge data consisted of BBC TV recordings of different genres [1]. The dataset was divided into three parts: training, development and evaluation. The training set consisted of 2193 episodes of 419 different shows with a total of 1600 hours of audio. This set included the BBC subtitles and metadata like speaker changes and time stamps. The organizers refined the metadata using a slightly supervised alignment [1]. We trained the UBM and the i-vector extractor on this set. The evaluation rules do not allowed using data from other sources to train i-vector extractors or PLDA.

The development set included 5 shows: "Doctor Who" (Show 1, 2 episodes), "Last of the summer" (Show 2, 6 episodes), "Springwatch" (Show 3, 3 episodes), "The Alan Clark Diaries" (Show 4, 6 episodes), "UEFA Euro2008" (Show 5, 2 episodes). The diarization task followed a longitudinal setup, i.e., the speakers of the current episode had to be linked to the speakers in previous episodes of the same show. The development set was used to evaluate the performance of our systems and to tune hyperparameters such as i-vector dimension, VBPLDA priors, etc. As this was the only set with reliable speaker labels, we used it to train the VBPLDA model posteriors. Also, we combined the development and train set to train VBPLDA with mixed and hidden speaker labels.

Finally, the evaluation set included 2 shows: "Celebrity masterchef" (Show 6, 11 episodes), "The culture show uncut" (Show 7, 8 episodes). A baseline clustering and VAD are provided with this dataset. The organizers ranked the participants submissions on this set.

Table 1: DER(%) on development set for different systems. GTS stands for ground truth segmentation, BICS denotes BIC segmentation, GTNS denotes ground truth Number of speakers, GTVAD denotes ground truth VAD, PLDA1 was trained on the dev. set and PLDA2 was trained on the train and dev. sets.

	Total	Show 1	Show 2	Show 3	Show 4	Show 5
Not linked						
GTS+GTNS+PLDA1	24.49	46.70	37.15	12.39	20.37	18.19
GTS+PLDA1 (cont. 1)	24.32	49.33	38.31	12.84	19.41	12.91
GTVAD+BICS+PLDA1 (cont. 4)	28.05	52.61	48.09	11.32	23.79	17.22
GTS+PLDA2 (primary)	**22.80**	**44.70**	**35.24**	13.44	**18.49**	**10.65**
GTVAD+BICS+PLDA2 (cont. 8)	27.41	53.76	47.01	**10.37**	22.76	17.39
Linked						
GTS+GTNS+PLDA1	28.07	47.25	44.19	12.85	24.44	24.83
GTS+PLDA1 (cont. 1)	27.72	49.81	46.84	13.11	24.38	13.67
GTVAD+BICS+PLDA1 (cont. 4)	31.05	53.54	54.35	11.65	29.12	17.70
GTS+PLDA2 (primary)	**25.96**	**45.65**	**40.51**	14.27	**24.70**	**11.14**
GTVAD+BICS+PLDA2 (cont. 8)	30.79	57.45	52.19	**10.80**	29.36	17.40

6.2. System configuration

The features of the system were 20 ETSI standard MFCC with short-time cepstral mean and variance normalization (CMVN) with a sliding window of 3 seconds. Our solutions to discriminate speech from music in TV shows are model based. As using models trained on other datasets was not allowed and the training set labels were not reliable to train new models, we decided to use the VAD provided by the organizers. On the development experiments, we used the ground truth VAD (GTVAD); and on the eval. set, we used the baseline VAD (BLVAD).

In the segmentation step, we divide each episode into short single speaker segments We used our segmentation based on BIC (BICS), and also the ones provided by the organizers, the ground truth segmentation (GTS) in development and the baseline segmentation (BLS) in the evaluation.

We trained a UBM of 256 Gaussians and an i-vector extractor with i-vectors of dimension 100 on the training set. We used the aligned metadata to define the segments where we compute the i-vector posteriors in the expectation step of the EM algorithm. In essence, for each short segment defined in the metadata, we computed an i-vector.

We applied centering, whitening and length normalization to the i-vectors [18]. The parameters needed for centering and whitening were trained also on the training set. For this step, speaker labels are not required.

We trained two VBPLDA models. The first PLDA (PLDA1) was trained only on the development set. We computed an i-vector for each segment defined in the ground truth clustering and then, trained PLDA with ground truth labels. We only selected the segments longer than 5 seconds belonging to speakers with more than 4 segments. Thus, we used 62 speakers with 1690 segments in total. The second PLDA (PLDA2) was trained on both the development and training sets. We used hidden speaker labels for the training set; the label posteriors and the PLDA model posteriors where obtained simultaneously as we iterate. We reduced the number of training segments by selecting segments longer than 5 seconds belonging to speakers with more than 8 segments– according to baseline clustering in the metadata. Thus, we kept 41995 segments. Then, we ran unsupervised VBPLDA assuming several number of speakers and selected the model that hypothesized 2500 speakers in the training set.

The speaker factor dimension was 50 for PLDA1 and 60 for PLDA2. Given the results in our previous works [12, 13], we put weak informative priors on the model parameters based on the average total variance of the data s_0^2. We assumed that the speaker space variance is 15% of s_0^2 and the channel space has 85% of s_0^2. Then, for α (prior of the inverse eigenvalues), we placed a wide prior with mode $1/(0.15s_0^2)$ by setting $a_\alpha = 2$ and $b_\alpha = 0.15s_0^2$. For $\mathbf{W}$, we used a Wishart prior with expectation $1/(0.85s_0^2)\mathbf{I}$ by setting $\nu_0 = 102$ and $\Psi_0 = 1/(0.85s_0^2\nu_0)\mathbf{I}$. Note that, for the Wishart prior to be proper, we need $\nu_0 > d$. We set $\tau_0 = 10/M$ where M is the number of hypothesized speakers.

7. RESULTS

Table 1 shows the Diarization Error Rate (DER) in the development set for different diarization systems. DER can be decomposed into missed speech, false alarm speech, and speaker detection error. In this case, the first two are not significant because we used the ground truth VAD (GTVAD) in this experiment. The false alarm was 0% and the missed speech was lower than 1% and it was due to undetected overlap between speakers. The first block of the table shows DER in a not-linked setup, i.e., we do not consider the errors made linking the speakers between episodes. Meanwhile, the sec-

Table 2: DER(%) on evaluation set for different systems. BLS stands for baseline segmentation, BICS denotes BIC segmentation, BLVAD denotes baseline VAD, PLDA1 was trained on the dev. set and PLDA2 was trained on the train and dev. sets.

	Total	Show 6	Show 7
Not linked			
BLS+PLDA1 (cont. 1)	44.67	48.57	38.34
BLVAD+BICS+PLDA1 (cont. 4)	42.77	46.65	36.48
BLS+PLDA2 (primary)	42.96	47.76	35.16
BLVAD+BICS+PLDA2 (cont. 8)	**39.86**	**43.18**	**34.47**
Best primary	40.20	44.59	**33.07**
Linked			
BLS+PLDA1 (cont. 1)	53.69	56.41	49.28
BLVAD+BICS+PLDA1 (cont. 4)	51.21	54.42	46.00
BLS+PLDA2 (primary)	50.48	55.27	42.69
BLVAD+BICS+PLDA2 (cont. 8)	**47.12**	**50.53**	**41.60**
Best primary	47.46	51.92	**40.21**

ond block shows the results for the linked setup.

The first row of each block shows results for the case where we cluster the i-vectors knowing the actual number of speakers. Meanwhile in the second row, the VBPLDA decides the number of speakers based on the variational lower bound. We point out that, most of the times, it did not detect the right number of speakers, however, it did not affect the DER too much. The results of PLDA1 and PLDA2 were close, PLDA2 was only about 1–2% absolute better than PLDA1. However, considering that PLDA1 was trained on the same data that we are evaluating, that could explain why adding more data to the PLDA training did not improve the DER. Using the BIC segmentation worsened around 5% absolute w.r.t the ground truth segmentation; it worsened 12% in the worst show but it improves 3.5% in Show 3. The DER of the linked condition only worsened about 3% absolute w.r.t to the not linked. The largest DER difference between linked and not linked happened for Shows 2 and 4 that were the ones with more episodes (6). As we process more and more episodes, the total number of speakers grows, which increases the chances of finding similar speakers and making linking errors.

We observe a large disparity in the DER between shows. One factor affecting performance is the quality of the speech. The show with the higher DER is Show 1 (Dr. Who), where speech is contaminated with background music and special effects. On the contrary, the show with the lowest DER has mostly clean speech. As commented above, another factor is the number of speakers in the show. For example, despite that Show 5 (UEFA championship) has crowd noise from the stadium, its DER is lower than the DER of Show 4, which is cleaner. However, Show 5 has 13 speakers per episode while Show 4 has about 30 speakers per episode.

Table 2 shows the DER of our systems on the evaluation data. In the last line, we include the result of the best primary system submitted by our competitors as reference. Our primary and contrastive 1 submissions used the baseline clustering provided by the organizers to create the initial segmentation. Meanwhile, contrastive 4 and 8 used the baseline VAD and our BIC segmentation. In this evaluation, we mainly focused on the clustering step. For this reason, we thought that the baseline segmentation might be better than ours but we were wrong. In fact, contrastive 8 slightly outperformed the best primary rival.

The baseline VAD had 6.1% of missed speech and 4% of false alarm speech. Thus, the speaker detection error is about 10% absolute lower than the DER in the table.

The eval. results are consistent with the dev. results. PLDA2 outperformed PLDA1. The difference between both is a bit larger than in the dev. set. The difference between the linked and not linked condition is significant, about 7% absolute. Once again, we think that this happens because of the high number of episodes and, therefore, speakers in the shows–186 and 300 speakers respectively.

8. CONCLUSIONS

We described the speaker diarization system that we proposed for the Multi-Genre Broadcast (MGB) Challenge at ASRU 2015. We proposed a system based on the i-vector paradigm. First, the system performs a segmentation step based on BIC. The segmentation finds points where the speakers' turns change and provides short segment containing a single speaker. Then, we compute an i-vector per speech segment and perform three clustering steps. A generative model based on Bayesian PLDA clusters the speakers. In this model, the speaker labels are latent variables that we optimize by variational Bayes iterations. The number of speakers in each episode was decided by maximizing the variational lower bound. After each clustering we merge some of the speech segments and re-compute the corresponding i-vectors. The i-vectors of the speakers of previous episodes were introduced into the clustering algorithm to link speakers between episodes.

We evaluated several system variants on the dev. and eval. data. Our best system (contrastive 8) obtained DER=30.79% on the dev. set and DER=47.12% on the eval. set. This system slightly outperformed the best primary system in the evaluation.

We observed that the quality of speech of the TV shows greatly affects performance. Shows with speech overlapped with music and special effects had much larger DER than shows with clean speech. Also, shows with larger number of episodes and speakers presented higher DER. We deduce that having more speakers increases the chances of finding speakers close in the i-vector space.

9. REFERENCES

[1] P Bell, Mark J. F. Gales, Thomas Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, Steve Renals, Óscar Saz, M. Wester, and P. C. Woodland, "The MGB Challenge: Evaluating Multi-Genre Broadcast Media Transcription," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015*, Scottsdale, Arizona, USA, Dec. 2015, IEEE.

[2] Sue E. Tranter and Douglas a. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1557–1565, Sept. 2006.

[3] Xavier Anguera Miro, S. Bozonnet, Nicholas Evans, Corinne Fredouille, G. Friedland, and Oriol Vinyals, "Speaker Diarization: A Review of Recent Research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, Feb. 2012.

[4] Scott S. Chen and P. Gopalakrishnam, "Speaker Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," in *DARPA Broadcast News Worshop*, 1998, pp. 127–132.

[5] Douglas A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2005*, Philadelphia, Pennsylvania, USA, Mar. 2005, vol. V, pp. 953–956, IEEE.

[6] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1448–1460, 2007.

[7] Fabio Castaldo, Daniele Colibro, Emanuele Dalmasso, Pietro Laface, and Claudio Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008*, Las Vegas, Nevada, USA, Mar. 2008, pp. 4133–4136, IEEE.

[8] Carlos Vaquero, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida, "Quality assessment for speaker diarization and its application in speaker characterization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 4, pp. 816–827, Apr. 2013.

[9] Douglas A. Reynolds, Patrick Kenny, and Fabio Castaldo, "A Study of New Approaches to Speaker Diarization," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association, Interspeech 2009*, Brighton, UK, Sept. 2009, pp. 1047–1050, ISCA.

[10] Fabio Valente, Petr Motlicek, and Deepu Vijayasenan, "Variational bayesian speaker diarization of meeting recordings," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2010*, Dallas, TX, USA, Mar. 2010, pp. 4954–4957, IEEE.

[11] Najim Dehak, Patrick Kenny, Redah Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-End Factor Analysis For Speaker Verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788 – 798, May 2011.

[12] Jesús Villalba and Eduardo Lleida, "Unsupervised Adaptation of PLDA by Using Variational Bayes Methods," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014*, Florence, Italy, May 2014, pp. 744–748, IEEE.

[13] Jesús Villalba and Eduardo Lleida, "Unsupervised Training of PLDA with Variational Bayes," in *Proceedings of Advances in Speech and Language Technologies for Iberian Languages, IberSpeech 2014*, Las Palmas de Gran Canaria, Spain, Nov. 2014, Lecture Notes in Artificial Intelligence, pp. 69–78, Springer International Publishing.

[14] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, Jan. 2000.

[15] Sandro Cumani, Ondrej Glembek, Niko Brummer, Edward De Villiers, and Pietro Laface, "Gender independent discriminative speaker recognition in i-vector space," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012*, Kyoto, Japan, Mar. 2012, pp. 4361–4364, IEEE.

[16] Patrick Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *Proceedings of Odyssey 2010 - The Speaker and Language Recognition Workshop*, Brno, Czech Republic, July 2010, ISCA.

[17] Jesús Villalba and Niko Brummer, "Towards Fully Bayesian Speaker Recognition: Integrating Out the Between-Speaker Covariance," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association, Interspeech 2011*, Florence, Italy, Aug. 2011, pp. 505–508, ISCA.

[18] Daniel Garcia-Romero and Carol Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association, Interspeech 2011*, Florence, Italy, Aug. 2011, pp. 249–252, ISCA.

[19] Christopher Bishop, "Variational principal components," in *Proceedings of the 9th International Conference on Artificial Neural Networks, ICANN 99*, Edinburgh, Scotland, Sept. 1999, IET, pp. 509–514.

[20] Christopher Bishop, *Pattern Recognition and Machine Learning*, Springer Science+Business Media, LLC, 2006.

[21] Kentaro Katahira, Kazuho Watanabe, and Masato Okada, "Deterministic annealing variant of variational Bayes method," *Journal of Physics: Conference Series International Workshop on Statistical-Mechanical Informatics 2007 (IW-SMI 2007)*, vol. 95, Jan. 2008.

[22] Jesús Villalba, "Unsupervised Adaptation of SPLDA," Tech. Rep., University of Zaragoza, Zaragoza (Spain), 2013.

[23] Michael R. Anderberg, *Cluster Analysis for Applications*, Academic Press, 1973.