SPEAKER DIARIZATION OF BROADCAST STREAMS USING TWO-STAGE CLUSTERING BASED ON I-VECTORS AND COSINE DISTANCE SCORING

Jan Silovsky and Jan Prazak

Institute of Information Technology and Electronics,Faculty of Mechatronics,
Technical University of Liberec,Czech Republic, {jan.silovsky,jan.prazak}@tul.cz

ABSTRACT

In this paper we present our system for speaker diarization of broadcast news based on recent advances in the speaker recognition field. In the system, speaker segments determined by the speaker change-point detector are represented by i-vectors and similarity of segments' speakers evaluated using cosine distance scoring. Linear discriminant analysis is employed to cope with intra-speaker variability. The experiments were carried out using the COST278 multilingual broadcast news database. We demonstrate improvement of the performance over the baseline system based on the Bayesian Information Criterion (BIC) and highlight significant impact of cepstral mean normalization. Finally, two-stage clustering employing BIC-based clustering to pre-cluster segments in the first stage is examined and showed to yield further performance improvement. The best performing configuration of our system achieved 52.4 % relative improvement of the speaker error rate over the baseline.

Index Terms— Speaker diarization, broadcast news, clustering, i-vectors

1. INTRODUCTION

An inherent part of a speaker diarization system is a clustering module. This paper presents our results from a study that investigated speaker clustering method applicable for broadcast stream audio processing. The method must be robust against large variety of recording conditions (often frequently changing) and transmission channels. The approach was initially proposed for the speaker recognition task [1] and proved to outperform the other state-of-the-art approaches in various NIST Speaker Recognition Evaluation (SRE) 2008 conditions, including the 10sec-10sec condition operating with short segments. The approach employs a simple factor analysis model to extract fixed-and low-dimensional representation of audio segments using so called i-vectors. Linear discriminant analysis is then applied in the i-vectors space to filter out the nuisance intra-speaker variability. Finally, similarity of segments' speakers is evaluated using cosine distance of the vectors. Utilization of approaches based on factor analysis in the task of speaker diarization was presented in [2] and application of clustering based on representation of speech segments by i-vectors was already reported in [3, 4]. Authors of [4] deal with diarization of summed (two-wire) telephone conversations and propose their solution with an assumption that recordings contain just two speakers. However, no such assumption is legitimate in broadcast domain which is of our interest and is handled in [3]. While author in [3] use average i-vector as estimated based on all i-vectors corresponding to the segments assigned to a cluster for its representation, we derive the i-vector representing the cluster as Maximum A Posterior (MAP) point estimate of factors in the total variability space based on summed sufficient statistics gather across all segments assigned to the cluster. Another difference compared to [3] consists in the way the initial segmentation is performed. Further, we present two stage clustering combining standard BIC-based approach with the approach based on i-vectors and highlight the impact of cepstral mean normalization in both stages of the clustering process.

Although we address particularly the clustering problem in this paper, we consider the assessment within the speaker diarization framework very useful as it reflects the ability to cope with possibly over-segmented speech or inexactly found segment boundaries leading to lower speaker purity of segments.

The research described in this paper was supported by the Technology Agency of the Czech Republic (project no. TA01011204) and by the Student Grant Scheme (SGS) at the Technical University of Liberec.

2. SPEAKER DIARIZATION SYSTEM

Our speaker diarization system consists of three basic modules. First, after feature vectors are extracted, speech activity detection (SAD) is applied. Then, speaker change points are detected by a speaker segmentation module. Finally, segments of the same speakers are clustered and speaker diarization is provided. All modules use standard Mel-frequency cepstral coefficient (MFCC) features.

The clustering module uses bottom-up clustering which is predominant approach for speaker clustering. First, a similarity measure between all pairs of speech segments is computed. Next, until the stopping criterion is met, the most similar pair of speech segments (clusters) is iteratively merged into a new cluster and the similarity measure between the new cluster and all remaining speech segments (clusters) is recomputed.

3. CLUSTERING METHODS

3.1. BIC-based clustering

The baseline system uses the Bayesian Information Criterion (BIC) as the similarity measure [5]. The BIC-based measure compares the BIC statistics of clusters g_1 and g_2 with the BIC statistic of a cluster g which is formed by merging of the cluster g_1 and the cluster g_2. We apply local BIC measure which is defined as

$$\begin{aligned}\Delta BIC(g_1, g_2) =&(N_1 + N_2)log\,|\mathbf{\Sigma}| - N_1 log\,|\mathbf{\Sigma}_1| \\ &- N_2 log\,|\mathbf{\Sigma}_2| - \alpha P\end{aligned} \quad (1)$$

where N is the number of frames, $\mathbf{\Sigma}$ is the full covariance matrix of the frames, α is a penalty weight and P is the penalty:

$$P = \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right) log(N_1 + N_2) \quad (2)$$

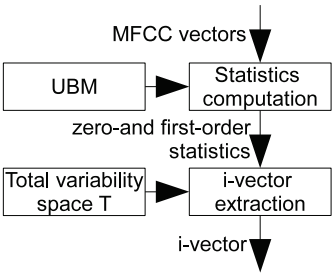


Fig. 1. The i-vector extraction process.

where d is the dimension of feature vectors.

In the clustering process, two clusters with the lowest ΔBIC value are merged together. If a minimal distance between any pair of clusters is higher than a certain threshold λ (typically zero), the stopping criterion is met.

The main advantage of the method is the small number of parameters that need to be estimated. No background models need to be estimated and the penalty weight α and the threshold λ represent the only parameters.

3.2. I-vectors extraction

We employ a simple factor analysis model to extract a fixed-dimensional representation of a segment of variable length as proposed by [1]. Let's assume a Gaussian Mixture Model (GMM) trained on data pooled from many speakers. This model is typically referred to as the Universal Background Model (UBM). The term *supervector* is used to refer to a high-dimensional vector obtained by concatenation of mean vectors of components of a GMM. Let $\boldsymbol{s}$ be a supervector representing a speech segment. In the i-vector concept, speaker-and segment-specific supervector for j'th segment of a speaker s is defined using a generative model as

$$\boldsymbol{s}_{s,j} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{x}_{s,j} \tag{3}$$

where $\boldsymbol{m}$ is a speaker-and segment-independent supervector (obtained from the UBM), the $\boldsymbol{T}$ is a rectangular matrix of low rank and the $\boldsymbol{x}_{s,j}$ is a random vector having standard normal distribution $\mathcal{N}[0, \boldsymbol{I}]$. The matrix $\boldsymbol{T}$ defines a total variability space and components of the vector $\boldsymbol{x}$ are corresponding factors. Following the terminology of [1] we refer to the vector $\boldsymbol{x}$ as the *i-vector*.

A projection from a sequence of feature vectors representing a speech segment to the i-vector space is provided by computation of a Maximum A Posterior (MAP) point estimate of the $\boldsymbol{x}_{s,j}$ based on zero-and first-order sufficient statistics gathered employing the UBM [6]. The process of the i-vector extraction is illustrated in Fig. 1. Considering the i-vector as a feature vector representing a speech segment, the factor analysis model (3) acts as part of the feature extraction process. Having a fixed-dimensional representation we can apply cosine distance scoring (CDS).

3.3. Cosine distance scoring

The cosine distance score CDS for segments (clusters) represented by i-vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is:

$$CDS = \frac{\boldsymbol{x}_1' \boldsymbol{x}_2}{\|\boldsymbol{x}_1\| \, \|\boldsymbol{x}_2\|}. \tag{4}$$

In the clustering process, the two clusters with the highest CDS value are merged together. If a maximum CDS value for any pair of clusters is lower than a certain threshold λ, estimated on the development data, the stopping criterion is met.

Now let $\boldsymbol{X}^{(g)} = \{\boldsymbol{x}^{(g)}_{1\ldots J^{(g)}}\}$ be a set of $J^{(g)}$ i-vectors representing segments assigned to a cluster g. Eq. 4 can be applied only for a pair of i-vectors, each representing one cluster in the trial. To obtain representation of a cluster by a single i-vector, sufficient statistics gathered employing the UBM for each segment assigned to the cluster are summed together and a MAP point estimate of the total factors extracted based on these summed statistics to form a single i-vector.

The LDA is employed to cope with the nuisance intra-speaker variability. The LDA defines an orthogonal projection matrix that maximizes between-classes variability and minimizes intra-classes variability. The projection matrix is estimated using a background data set. In our case, each class is formed by all segments of a single speaker in an audio stream.

3.4. Two-stage clustering

Two-stage clustering scenario was successfully applied in speaker diarization systems dealing with meeting, lecture or broadcast data [7]. Our motivation stems from a hypothesis that the MAP point estimate of the total factors (i-vectors) for segments of short duration cannot be estimated reliably which may harm the clustering process particularly at early phases. To cope with the problem we employ two-stage clustering scenario. We have already experienced good results obtained with the two-stage approach in the system employing Probabilistic Linear Discriminant Analysis (PLDA) modeling of i-vectors [8]. In the first stage, we use BIC-based clustering with zero value of the stopping threshold λ and a value of the BIC penalty weight α set so as to under-cluster the segments and thus reduce the number of very short segments (shorter than 0.5 s). In the next stage, the clustering approach employing i-vectors is applied.

4. EXPERIMENTS AND RESULTS

4.1. Datasets

Experiments were carried out using the COST278 multilingual pan-European broadcast news database [9]. The database comprises broadcast news recordings in 9 languages. Authors of the database divided the data for each language into a training set (containing about two hours) and a test set (containing about one hour).

We split the data into three datasets. The first set contained all COST278 Croatian, Czech, Hungarian, Portuguese and Slovak training data giving in total 11.5 hours of audio. This set was used for training of the UBM, estimation of the total variability space and LDA projection matrix. The second set, consisting of 13 shows of various lengths (in the range from 8.5 to 53.8 minutes) drawn also from the COST278 training data and giving in total 5.89 hours, was used as the development set for tuning of system parameters. Particularly for estimation of segmentation and clustering stopping thresholds. Finally, the third set was used as the test set in our experiments. The set consisted of 15 shows of various lengths (in the range from 4.1 to 53.2 minutes) drawn from the COST278 test data and giving in total 6.34 hours. The development and test data comprised of 5 languages: Belgian Dutch, Czech, Hungarian, Slovak and Slovenian. The streams in COST278 corpus contain also commercials which are not annotated. The commercials were thus removed from the streams used in development and test sets.

4.2. Training data utilization

The UBM with 1024 components was trained using the data from 1007 speakers (2530 segments, 11.5 hours). The total variability space was estimated using a subset of the UBM training data resulting from the condition of minimal length of a segment of 3 seconds and using at most eight segments per speaker. This resulted in 2050 segments (10.2 hours) from 909 speakers. The LDA projection matrix was estimated using the data from speakers for which at least three segments of minimal length of 3 seconds are available, in total 1528 segments (7.5 hours) from 280 speakers were used. The average length of segments used in training is 17.8 s.

4.3. Evaluation metrics

Performance of diarization systems is usually evaluated by the Diarization Error Rate (DER) as the primary metric [10]. The DER can be decomposed as $DER = SPKE + FA + MISS$, where the SPKE represents the speaker error rate, the FA is the speech false alarm error rate and the MISS is the missed speech error rate. The SPKE reflects the amount of speech data that is attributed to a wrong speaker given the optimum speaker mapping between a system output and a reference diarization. Because all our evaluated systems share the same SAD and speaker segmentation modules, we use the SPKE as the primary metric. Likewise in [10], a forgiveness collar of 0.25 s (both + and -) was not scored around each boundary.

4.4. Features extraction

All components of the system use classic Mel-frequency cepstral coefficient (MFCC) features. We used 25 ms window and 10 ms window shift. The segmentation and clustering modules use feature vectors formed from 30 and 12 static MFCCs respectively. Cepstral mean normalization (CMN) was not employed within the segmentation phase and application of CMN for clustering will be discussed in next sections. In case CMN was employed, it was applied for a center frame within a sliding window of length of 400 frames.

4.5. Baseline system results

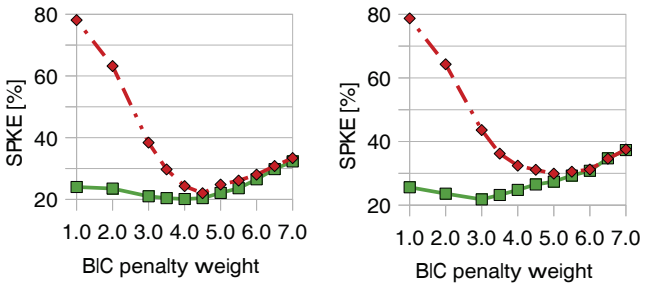
The SAD module achieved FA of 0.8 % and MISS of 3.2 % on the test set. We found that higher value of the MISS is caused by inaccuracy of reference annotations. In fact, we have observed that our speech activity detection was able to detect the boundaries of speech segments precisely. The average length of speech segments after segmentation was 3.6 s.

First, performance for different values of the BIC penalty weight α was evaluated. We found that systems using a non-zero value of the stopping threshold λ estimated on the development data yielded better performance than the systems operating with zero value of the threshold, see Fig. 2. The best performance on the development set was obtained for the system using penalty weight of 4.0. On the test set, the system achieved SPKE of 24.8 % which corresponds to the DER of 28.8 %. These results are considered as baseline. Further, we found the application of CMN to cause significant performance degradation, particularly for higher values of the BIC penalty weight. Fig. 3 illustrates the effect of CMN application for the systems operating with non-zero stopping threshold values.

4.6. Cosine distance scoring system

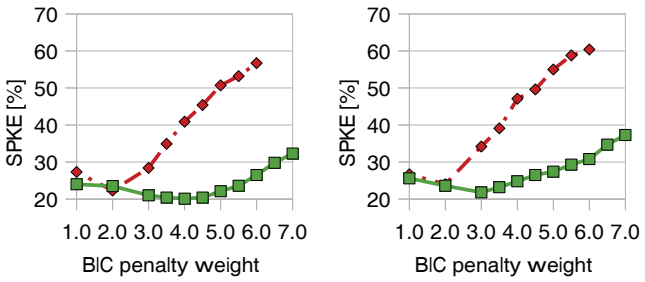
Various configurations of the system employing cosine distance scoring, differing in the number of Gaussians in the UBM, dimensions of the total variability space and LDA dimension reductions, were examined. The best results were obtained for setups using the UBM with 256 Gaussians to extract the sufficient statistics. Fig. 4(a) shows the effect of different LDA dimension reductions for systems operating with total variability spaces of dimensions of 300 and 400. The zero LDA dimension stands for no application of LDA in Fig. 4. Fig. 4(b) shows the effect of CMN for the system using 400-dimensional i-vectors. We remark that application of CMN requires utilization of the UBM, the total variability space and the LDA projection trained using the mean normalized data.

Systems based on CDS provide performance similar to the baseline system when LDA is not applied. Compared to the baseline, the SPKE was slightly reduced from 24.8 % to 24.1 % (relatively by 2.8 %) by the system using 300-dimensional i-vectors. When the nuisance variability is projected out by virtue of LDA, the performance is significantly improved. The best SPKE of 16.9 % (31.9 % relative improvement) was achieved by system employing 400-dimensional i-vectors and the LDA dimensional reduction to 200. Concerning the effect of CMN, no significant impact was observed compared to the severe impact in the case of the baseline BIC-based system. For most of evaluated setups (including those not enclosed in Fig. 4(b)), application of CMN slightly improves performance.



(a) devel set (b) test set

Fig. 2. Speaker error rate evaluated on (a) the development set and (b) the test set for BIC-based clustering with zero (dashed line) and estimated (solid line) stopping threshold.

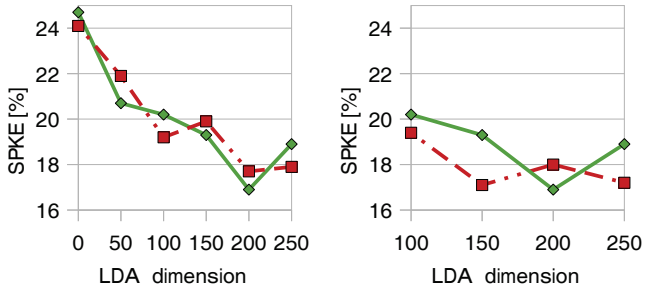


(a) devel set (b) test set

Fig. 3. Results for BIC-based clustering with (dashed line) and without (solid line) application of CMN.

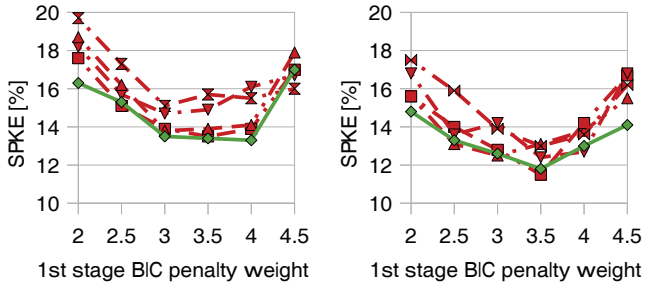
4.7. Two-stage clustering results

Significant impact of the BIC penalty weight, applied in the first BIC-based pass of the two-stage clustering approach, was observed. Fig. 5(a) shows results achieved without application of CMN at the



(a) LDA dimension effect (b) CMN effect

Fig. 4. (a) Results for CDS-based clustering with 300-dimensional (dashed line) and 400-dimensional (solid line) i-vectors. (b) Effect of CMN (dashed line) for system with 400-dimensional i-vectors.



(a) without CMN at 2nd stage (b) with CMN at 2nd stage

Fig. 5. (a) Effect of the BIC penalty weight applied in the first stage of two-stage clustering scenario for various setups of the CDS-based clustering used in the second stage. Solid line highlights the best setup. (b) The same situation with CMN applied at the second CDS-based stage.

second clustering stage based on i-vectors and CDS. The best performing setup is highlighted by the solid line. All evaluated setups yield the best performance for the BIC penalty weight in the range from 3.0 to 4.0. Higher value of the BIC penalty weight invokes less penalization for long segments and thus leads to lower under-clustering degree. Fig. 5(b) shows the results achieved with application of CMN in the second clustering stage. We remark that CMN is not applied at the first stage in either case. Application of CMN remarkably improves performance of all systems compared to the one-stage clustering scenario. Also the optimal value of the BIC penalty weight applied in the first stage seems to be easier to determine as all systems provide best performance for the penalty weight of 3.5. The best performing system employing 400-dimensional i-vectors and LDA dimensional reduction to 200 achieved SPKE of 11.8 % (52.4 % rel. impr.).

Our reasoning for different effect of CMN in case of one-stage and two-stage clustering is following. The first BIC-based clustering stage is less prone to fail because of short duration of segments being clustered. Moreover, as CMN is not applied, the BIC-based stage aims to cluster the adjacent segments and thus suppress the over-segmentation. At the second clustering stage, the number of short segments is reduced and thus more reliable MAP estimates of i-vectors are available. Further, CMN applied at the second stage facilitates clustering of remote segments of the same speakers acquired under distinct conditions.

5. CONCLUSIONS

In this paper, we have described our speaker diarization system utilizing clustering approach based on i-vectors extraction and cosine distance scoring. Linear discriminant analysis is employed to cope with the intra-speaker variability. While the system without LDA applied provide performance similar to the baseline system, when the nuisance variability is projected out by virtue of LDA, the performance is significantly improved. Next, application of cepstral mean normalization yields further improvement of performance. The speaker error rate was reduced relatively by 31.9 % in the one-stage clustering scenario. In the two-stage clustering scenario, speaker error rate was further reduced. The best system achieved SPKE of 11.8 % which corresponds to 52.4 % relative improvement over the baseline error rate of 24.8 %.

One of drawbacks of the presented speaker clustering technique is higher computational cost. The real-time factor of the processing time for the baseline system is 0.04 while for the best performing two-stage system it is 0.07 (measured on a machine with Intel Core i7@2.66GHz). Another drawback is the need of extensive database containing several recordings of many speakers for estimation of the LDA projection matrix.

6. REFERENCES

[1] 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, May 2011.

[2] 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, December 2010.

[3] J. Franco-Pedroso, I. Lopez-Moreno, D.T. Toledano, and J. Gonzalez-Rodriguez, "ATVS-UAM system description for the audio segmentation and speaker diarization Albayzin 2010 evaluation," in *FALA 2010*, November 2010, pp. 415–417.

[4] S. Shum, N. Dehak, E. Chuangsuwanich, D.A. Reynolds, and J.R. Glass, "Exploiting intra-conversation variability for speaker diarization," in *Interspeech'11*, Florence, Italy, August 2011, pp. 945–948.

[5] S.S. Chen and P.S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proceedings 1998 DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 127–132.

[6] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Processing*, vol. 13, May 2005.

[7] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain, "Combining speaker identification and BIC for speaker diarization," in *Interspeech'05, ISCA*, Lisbon, September 2005.

[8] J. Silovsky, J. Prazak, P. Cerva, J. Zdansky, and J. Nouza, "PLDA-based clustering for speaker diarization of broadcast streams," in *Interspeech'11*, Florence, Italy, August 2011, pp. 2909–2912.

[9] An Vandecatseye et al., "The COST278 pan-European broadcast news database," 2004, pp. 873–876.

[10] NIST, "The 2009 (RT-09) rich transcription meeting recognition evaluation plan," 2009.