Analysis of I-Vector framework for Speaker Identification in TV-shows

Corinne Fredouille[1], *Delphine Charlet*[2]

[1]University of Avignon, CERI/LIA, France
[2]Orange Labs, Lannion, France
corinne.fredouille@univ-avignon.fr,delphine.charlet@orange.com

Abstract

Inspired from the Joint Factor Analysis, the I-vector-based analysis has become the most popular and state-of-the-art framework for the speaker verification task. Mainly applied within the NIST/SRE evaluation campaigns, many studies have been proposed to improve more and more performance of speaker verification systems. Nevertheless, while the i-vector framework has been used in other speech processing fields like language recognition, a very few studies have been reported for the speaker identification task on TV shows. This work was done in the REPERE challenge context, focused on the people recognition task in multimodal conditions (audio, video, text) from TV show corpora. Moreover, the challenge participants are invited for providing systems for monomodal tasks, like speaker identification. The application of the i-vector framework is investigatedthrough different points of views: (1) some of the i-vector based approaches are compared, (2) a specific i-vector extraction protocol is proposed in order to deal with widely varying amounts of training data among speaker population, (3) the joint use of both speaker diarization and identification is finally analyzed. Based on a 533 speaker dictionary, this joint system wins the monomodal speaker identification task of the 2014 REPERE challenge.

Index Terms: speaker identification, i-vector, REPERE challenge, TV shows

10.21437/Interspeech.2014-15

1. Introduction

The REPERE challenge is a project funded by the French Research Agency (ANR) and the French defense procurement agency (DGA) [12]. Its aim is to support research on people recognition in multimodal conditions. Since 2012, annual evaluation campaigns were organized to evaluate automatic systems developed by the research consortia involved. Competitive systems have to answer the following questions : "who is speaking ?", "who is present in the video ?", "what names are cited ?", "what names are displayed ?", by using, alone or combined, information issued from the audio, video and text streams. Each question can be handled by participants through a supervised or unsupervised mode (with or without biometric models). In addition, more "basic" systems are also evaluated, involving speech transcription, speaker diarization, named entity detection, OCR recognition, face segmentation, etc. A large effort was made by the REPERE challenge organizers to produce relevant and rather large annotated corpora progressively for both development of systems and their evaluation for the annual campaigns. These corpora cover different TV shows issued from two French TV channels (BFM TV and LCP), including news and debates. This paper is dedicated to the monomodal supervised speaker identification task within the REPERE challenge. More precisely, it is proposed to observe and discuss the behavior of i-vector framework-based systems for this typical task in TV shows. Although the i-vector framework has been largely applied for the speaker verification task in the specific NIST/SRE evaluation context or in other speech processing fields like language recognition [10] or speaker attribution [15], a very few studies have been reported in the literature concerning speaker identification in TV shows. However, this context implies some particularities non trivial to deal with like the widely varying amount of data per speaker present in the shows according to their role (recurrent anchor speakers, popular or punctual speakers, etc.) necessary for the speaker modeling, the granularity of segments implied in the identification decision while processing an entire TV show, generally issued from a preliminary speaker diarization step, and finally the coverage of dictionary used by the speaker identification system and its impact on an open-set identification task, which is closer from real life applications.

2. I-vector-based approaches

In Joint Factor Analysis (JFA) [3], speaker model training relies on the separate estimate of both a speaker and a channel/session subspaces in order to take into account the channel/session variability explicitly and to be able to compensate it in the context of Gaussian Mixture modeling. By observing that the channel/session subspace could retain some speaker information, Dehak et al. [8] have proposed a simpler and very powerful modeling paradigm based on a single total variability space, making no distinction between speaker and channel/session information. Here, the speaker- and channel-dependent GMM supervector, M, issued from the concatenation of speaker GMM means can be defined as :

$$M = m + Tw \quad (1)$$

where m is the mean supervector issued from the Universal Background Model (UBM) representing the speaker- and session/channel-independent information, the low-rank matrix T defines the total variability space, and w represents the speaker- and session/channel-dependent factors in the total variability space, also called i-vectors.

Research work around the i-vector framework has been fruitful, with the aim of enhancing their efficiency as well as of taking advantage, in terms of complexity reduction, of the low dimension spaces involved compared with the JFA framework. First of all, different scoring approaches have been proposed in order to make decision from i-vectors extracted from both training and testing utterances. Among these scoring approaches, we can differentiate the simple Cosine Distance Scoring (CDS)[8] and derived distances to deal with score normalization [4] from more advanced Gaussian-based scoring approaches like two-covariance scoring [7], Mahalanobis scoring [6], and Gaussian

Probabilistic Linear Discriminant Analysis (PLDA)[2], or from Heavy Tailed PLDA as well, in which Gaussian priors have been replaced with Student's t distributions [5].

As mentioned above, no distinction is made within the total variability space between speaker and channel/session information. Consequently, a simple CDS-based system may take benefit of channel/session compensation approaches to deal with. Different channel compensations have been investigated in the literature: Within-Class covariance normalization (WCCN), Linear Discriminant Analysis (LDA)[8], or nuisance attribute projection (NAP) [1], used separately or combined.

Finally, different studies have been focused on i-vector normalization techniques in order to ensure the expected Gaussianity of i-vector distribution, while Gaussian-based scoring approaches are implied. Among the i-vector normalization proposed in the literature, we can cite the straightforward length normalization (division of the i-vectors by their Euclidian norm) proposed in [9] or more advanced ones like variance-spectra based-normalization techniques proposed in [11], named Eigen-Factors Radial (EFR) or Spherical Nuisance Normalization (SphNorm).

3. The speaker identification system

In this paper, the authors propose a typical i-vector framework-based system for the supervised monomodal speaker identification task of the 2014 REPERE challenge. In order to deal with TV shows, the latter is joint to a speaker diarization system, which provides speech segments belonging to speakers as well as speaker clusters, both on which the speaker identification can be applied. Next sections will described these processes.

3.1. Speaker diarization

The diarization system used in this work is the one presented in [14]. It is a sequential processing using firstly Bayesian Information Criterion and then Cross-likelihood Criterion, with special attention paid for overlapped speech for TV-debates, where the amount of overlapped speech is significant. For these shows, overlapped speech segments are first detected and discarded from the clustering process, and then reassigned to the 2 nearest speakers, in terms of temporal distance between speech segments. For news shows, overlapped speech is considered negligible, and this process is not applied.

3.2. I-vector based speaker identification system

The speaker identification system used in this paper, and notably, the i-vector-based framework, relies on the ALIZE v3.0 toolkit [16]. The following sections will describe which and how i-vector-based tools are applied on TV show corpora within the REPERE challenge.

3.2.1. Total variability space and i-vector extraction

19 LFCC augmented with their delta coefficients, the delta energy, and 11 double delta coefficients are used for the feature extraction. Features are then normalized, file by file, by applying a cepstral mean subtraction and variance normalization. The i-vector extraction relies on a 200 dimension total variability space estimated from about 1200 speakers and 7500 sessions. The Universal Background Model (UBM) is gender independent, represented by a 512 component Gaussian Mixture Model. It is learnt on about 200h of speech.

As reported above, the training corpus available for the 2014 evaluation campaign of the REPERE challenge is quite large, including 47h of speech issued from TV debates and news. This context is very different from the NIST evaluation campaigns, for which speaker models are estimated on utterances varying from 10s to 2mn30, depending on the targeted condition. Here, training data for a speaker may vary from a very few seconds to more than 2 hours. Since i-vector framework based systems reach very high performance on 2mn30 training condition of the last NIST evaluation campaign, training i-vector extraction for a given speaker was applied, in this paper, following the couple of rules : (1) an i-vector is extracted only if a minimum 30s long training data are available, (2) if training data for a given speaker is longer than 2mn30, a set of i-vectors was extracted, each of them on the basis of 2mn30 duration, the last one having to respect the rule (1).

Considering now the testing i-vectors, the extraction relies on the output of the speaker diarization system. Two extraction paradigms were investigated : (1) either considering each speech segment issued from the diarization output individually without taking into account the cluster information and extracting a corresponding segment-based i-vector, or (2) considering all the speech segments gathered in a cluster and extracting a cluster-based i-vector. In the former, very short speech segments (a few seconds) may be encountered, which may impact on the i-vector robustness. Nevertheless, it could permit to overcome some clustering errors produced by the speaker diarization system. On the other side, the decision per cluster permits to handle overlap speech since a segment may be processed twice if it is attributed to a couple of clusters during the speaker diarization process.

3.2.2. Scoring and normalization

Cosine Distance Scoring (CDS) and Probabilistic Linear Discriminant Analysis (PLDA) scoring approaches were used for experimental comparisons. The former was applied alone or associated with Within-Class covariance normalization for session/channel compensation. In the same manner, PLDA was used alone or combined with either the EigenFactors Radial (EFR) or Spherical Nuisance Normalization (SphNorm) for i-vector normalization. Channel rank is equal to the dimension of the i-vectors (200) whereas speaker rank is equal to 100. For WCCN and PLDA, matrix estimates were based on 571 speakers and 4384 sessions.

During scoring, two different manners of combining multiple training i-vectors per speaker, where needed, were investigated, either by averaging scores obtained by each individual training i-vectors (when compared with a single testing i-vector), or by taking the maximum score over the ones available. Finally, all the scores are t-normed, notably for the open-set speaker identification task.

4. Experimental protocols

Experiments conducted in this paper were based on the framework of the REPERE Challenge 2014 evaluation campaign. They were all based on three different corpora available in the challenge : a 47h-long training corpus, named FullTrain in the rest of the paper, a 3h-long development set named Phase2-dev and a 10h-long test set, named Phase2-test, available for the final 2014 evaluation only. These corpora were annotated in terms of speaking and visible persons, in a continuous way for the speech stream and in a discrete way for the video stream, leading to an annotated video keyframes every 10s in average.

	FullTrain	Phase2_dev	Phase2_test
Corpus size	47h	3h	10h
# speak. in dict./# identifiable speak. in ref. (% speakers covered)	289/788(36.7)	55/103(53.4)	108/270(40)
# keyframes covered by the dict./# identifiable keyframes in ref.	14209/17681(80.4)	703/943(74.5)	2829/3831(73.8)

Table 1: Information relating to the REPERE datasets, FullTrain, Phase2_dev, Phase2_test, including their size (in hours), the coverage of both speakers and keyframes according to the dictionary and references (numbers and %).

If the speaker identification was applied in a continuous way, the evaluation was based on the keyframe level, following the REPERE challenge evaluation protocol [12].

While the Phase2_dev and Phase2_test corpora were used for testing only, the FullTrain corpus was implied both for speaker model training and testing by applying a Leave-One-Out paradigm (while a TV show is tested, the training i-vectors possibly available for the set of speakers present in the show are automatically removed). These corpora comprise 14209, 703, and 2829 testing keyframes respectively for the closed set speaker identification task (speakers present in keyframes are known by the identification system i.e. they are in the speaker model dictionary of the system) and 17681, 943, and 3831 testing keyframes respectively for the open-set task (speakers can be unknown).

Given that the REPERE challenge focuses mainly on news and political debates, the speaker model dictionary involved in the identification system includes 533 journalists and politicians. It was built according to the minimum duration rules described in section 3.2.1. These people were present in the FullTrain corpus as well as in some additional, but similar, TV or radio corpora or in some video excerpts available on the web. Table 1 provides, per corpus, some information about the dictionary coverage according to the population of identifiable people[1] present in the corpora as well as to the set of associated testing keyframes. It can be observed that, even if the dictionary covers, in average, 43% of identifiable people only, it covers, in average, 76% of associated keyframes.

Finally, the estimate of the total variability space, the UBM model or the PLDA/WCCN approaches were conducted on some separate French corpora, based on radio and TV show recordings.

5. Results and discussions

5.1. I-vector-based system

Table 2 provides performance, in terms of Correct Identification Rate, for different configurations of the speaker identification system. Since the behavior of techniques involved in the speaker identification system is analyzed in this section, evaluation is concerned here with a closed-set speaker identification task, i.e. among the population of speakers present in the TV shows, only those available in the system dictionary, and their associated tested keyframes, are considered in the performance evaluation. Moreover, the speaker identification process is applied at the cluster level, which means that the i-vector extraction as well as the decision are made for each cluster issued from the speaker diarization system. Finally, in the context of multiple training i-vectors per speaker (see section 3.2.2), the selection of the maximum score is used. Next, both methods are compared.

[1] Anonymous people, who had been annotated in the corpora, are not considered here.

Approach	FullTrain	Phase2_dev	Phase2_test
GMM modeling	81.4	77.9	88.8
Cosine distance	89	**88.6**	91.1
Cosine+WCCN	**90.3**	88.2	**91.4**
EFR+PLDA	89.8	86.9	90.2
SphNorm+PLDA	88.7	86	85.9
Ref. segmentation Cosine+WCCN	97.3	93.7	97.5

Table 2: Performance, in terms of Correct Identification rates (CIR in %), for the closed-set speaker identification task, depending on different approaches used for speaker modeling (GMM- or i-vector-based), and for i-vector techniques.

First of all, the comparison between a classical GMM-based system[2] and different i-vector configurations, shows, as expected, the supremacy of the latter, notably on the FullTrain and Phase2_dev datasets with about 9% of absolute gain.

Considering now the different configurations relating to the i-vector framework, the effectiveness of the simple Cosine distance, outlined in the literature, is quite confirmed here, on the three testing corpora. The benefit of the WCCN channel/compensation technique is rather minor, except for the FullTrain corpus with a 1.3% gain. Performance reached by the PLDA, combined either with EigenFactors Radial (EFR) or Spherical Nuisance Normalisation (SphNorm), is worse whatever the corpus observed. It seems that the choice of additional data required for PLDA, EFR, SphNorm (even WCCN-based technique) seems to be still more critical than for the NIST evaluation campaign. It is assumed that very varying duration conditions within training data or between training and testing data might be a first reason to this lower performance compared with the simple Cosine distance.

Finally, the power of the Cosine distance is confirmed through the identification rates obtained when applied on the reference speaker segmentation (last row in table 2) instead of the one issued from the automatic speaker diarization system. Here, we can observe very high rates, especially for the FullTrain and Phase2_test corpora, with 97.3% and 97.5% CIR respectively.

Based on the Cosine distance+WCCN technique configuration, table 3 compares performance of the couple of approaches proposed to deal with multiple training i-vectors per speakers. By selecting the score of the best i-vectors for a given speaker according to the set of ones available, instead of averaging all the scores, achieves best performance for all the testing corpora. The selection of the best training i-vector should probably permit to reduce the potential mismatch between training and testing data session.

[2] relying on the same UBM as used in the i-vector framework and a classical MAP adaptation for the speaker models.

Decision approach	FullTrain	Phase2_dev	Phase2_test
Score mean	89	87.8	90.9
Maximum score	**90.3**	**88.2**	**91.4**

Table 3: Performance in terms of Correct Identification rates (CIR in %) for the closed-set speaker identification task, depending on two different approaches to deal with multiple training i-vectors per speaker (Cosine+WCCN at the cluster level).

	FullTrain	Phase2_dev	Phase2_test
confusion	6.3	6.5	7.4
miss	12.0	14.1	11.1
false alarm	6.8	5.4	5.5
Precision	85.4	86.1	85.6
Recall	76.6	74.2	76.3
Fmeasure	80.7	79.7	80.7

Table 4: Performance of open-set speaker identification, scoring at the segment-level

5.2. Open-set speaker identification

In previous sections, evaluations are conducted in the closed-set speaker identification framework, so as to strictly focus on the modelling capacities of the i-vector. Here, evaluations are extended to the open-set speaker identification which is the realistic framework for TV-shows, where it is impossible to have a model for all the speakers who appear. The task is then to identify the speakers in the dictionary and to reject the speakers out of the dictionary. Thus, the rejection process is simply based on a threshold on the identification score, and the evaluation translates the ability of the score not only to rank correctly the hypotheses (the true speaker must have the 1-best score), but also the ability to reject unknown speakers (the unknown speakers must have a score below a certain threshold).

Here, the i-vector configuration, based on the Cosine distance combined with the WCCN technique and the maximum score selection for the multiple training i-vectors, is used. Furthermore, a T-norm-based normalization of identification scores is performed to facilitate the threshold definition and the rejection process. Finally, both segment-based and cluster-based configurations are involved in the following sections.

In this open-set paradigm, in addition to confusion errors, rejection of in-dictionary speakers and false acceptance of unknown speakers can occur. Performance can also be estimated in precision (percentage of correctly identified occurrences relatively to the set of identified occurrences) and recall (percentage of correctly identified occurrences relatively to the set of in-dictionary speakers), along with their harmonic mean, the F-measure.

Table 4 details performance obtained in the open-set paradigm for the three datasets. Evaluation is performed with the threshold that gives the maximal F-measure on the FullTrain set. This threshold is then applied on the other datasets when scoring at the segment level. Results show that rather good performance is achieved when it comes to reject unknown speakers, and that performance is quite stable across corpora.

5.3. Impact of speaker diarization

In order to evaluate the impact of speaker diarization, 4 constrastive experiments are performed: scoring at the segment level with reference segmentation (ref-seg), at the cluster level with the reference clustering (ref-clus), at the segment level with automatic segmentation (auto-seg) and at the cluster level with the automatic clustering (auto-clus). Results are evaluated in the open-set speaker identification framework, with the threshold that maximizes F-measure on FullTrain set and shown in table 5. The automatic diarization system provides a Diarzation Error Rate of 12.5% on FullTrain, 13.6% on phase2_dev and 14.2% on phase2_test, when including overlapping speech in the standard NIST evaluation metrics.

	FullTrain	Phase2_dev	Phase2_test
ref-seg	81.1	82.1	83.5
	(88.0,75.2)	(88.4,76.6)	(89.4,78.3)
ref-clus	85.5	86.7	88.0
	(89.0,82.2)	(92.2,81.8)	(90.9, 85.3)
auto-seg	80.7	79.7	80.7
	(85.4,76.6)	(86.1,74.2)	(85.6,76.3)
auto-clus	80.0	79.3	80.1
	(81.8, 78.2)	(83.5,75.5)	(81.6,78.7)

Table 5: impact of speaker diarization for open-set speaker identification: Fmeasure (Precision, Recall)

It can be seen that, when working on reference (i.e. perfect) segmentation and clustering, scoring at the cluster level helps a lot, improving both precision and recall, compared with scoring at the segment level, which is not surprising, as more data are available to identify speakers.

When it comes to automatic segmentation, conclusions are quite different. On the overall performance given by the F-measure, results are quite comparable between segment-level scoring and cluster-level scoring, because recall rate is increased by the cluster-level scoring whereas the precision rate is decreased. Indeed, accumulating more data in the cluster for the scoring can help to retrieve short segments and increase the recall, but on the other hand, the errors of purity in the clustering can lead to mix multiple voices in a same cluster, and decrease the precision rate. Hence, I-vector modelling which already achieves very good performance at the segment level is very sensitive to purity errors at the cluster level.

6. Conclusion and future work

This paper presents the application of i-vector technique to the specific task of speaker identification on TV shows in the context of the challenge REPERE. Experimental results have shown the powerful of the Cosine distance, associated with the Within Class Covariance normalization technique, compared with the PLDA techniques for the closed-set and open-set identification task. Moreover, regarding the impact of the preliminary speaker diarization task, performance analysis has highlighted the sensitivity of a speaker cluster-based approach to the cluster purity. Further work will have to investigate more the impact of duration mismatch between training and testing data, since the TV-show context may emphasize still more this issue compared with the NIST evaluation context, which is the most studied in the literature.

7. Acknowledgements

This work was funded by the French Research Program ANR Project PERCOL in REPERE Challenge (ANR 2010-CORD-102). The authors would like to thank Anthony Larcher and Pierre-Michel Bousquet for their advices on the use of the ALIZE 3.0 toolkit and i-vector framework.

8. References

[1] Campbell, W. M. , Sturim, D. E. and Reynolds, D. and Solomonoff, A.,"SVM based speaker verification using a GMM super vector kernel and NAP variability compensation", in Proc. of ICASSP, Toulouse, 2006.

[2] Prince, S. J. D., "Probabilistic linear discriminant analysis for inferences about identity", in Proc. of International Conference on Computer Vision (ICCV), Rio de Janeiro, Brazil, 2007.

[3] Kenny, P. and Ouellet, P. and Dehak, N, and Gupta, V. and Dumouchel, P., "A study of interspeaker variability in speaker verification", in IEEE Transactions on Audio, Speech, Language processing, Vol. 16(5), July 2008.

[4] Dehak, N, and Dehak, R. and Glass, J. and Reynolds, D. and Kenny, P., "Cosine similarity scoring without score normalization techniques", in Proc. of Odyssey - The Speaker and Language Recognition Workshop, pp. 71-75. Brno, Czech Republic, 2010.

[5] Kenny, P., "Bayesian speaker verification with heavy-tailed priors", in Proc. of Odyssey - The Speaker and Language Recognition Workshop, Czech Republic, 2010.

[6] Bousquet, P.-M. and Matrouf, D. and Bonastre, J.-F., "Intersession compensation ans coring methods in the i-vector space for speaker recognition", in Proc. of International conference on Speech Communication and Technology, 2011.

[7] Brummer, N., and Villalba, J. and Lleida, E., "Fully bayesian likelihood ratio vs i-vector length normalization in speaker recognition systems", in NIST SRE analysis workshop, 2011.

[8] Dehak, N, and Kenny, P. and Dehak, R. and Dumouchel, P. and Ouellet, P., "Front-end factor analysis for speaker verification", in IEEE Transactions on Audio, Speech, Language processing, Vol. 19(4), May 2011.

[9] Garcia-Romero, D. and Espy-Wilson, C. Y., "Analysis of I-vector Length Normalization in Speaker Recognition Systems", in Proc. of Interspeech, Florence, Italy, 2011, pp. 249-252.

[10] Martnez, D. and Pichot, O. and Burget, L. and Glembek, O. and Matejka, P., "Language Recognition in iVectors Space", in Proc. of Interspeech, Florence, Italy, 2011.

[11] Bousquet, P. and Larcher, A. and Matrouf, D. and Bonastre, J.-F. and, Pichot, O., "Variance-Spectra based Normalization for I-vector Standard and Probabilistic Linear Discriminant Analysis", In Proc. of Odyssey - The Speaker and Language Recognition Workshop, Singapur, 2012.

[12] Giraudel, A. and Carr, M. and Mapelli, V. and Kahn, J. and, Galibert, O. and Quintard, L., "The REPERE Corpus : a multimodal corpus for person recognition", In Proc. of LREC, Istanbul, 2012.

[13] Sarkar, A. K. and Matrouf, D. and Bousquet, P.-M. and Bonastre, J.-F., "Study of the Effect of I-vector Modeling on Short and Mismatch Utterance Duration for Speaker Verification", in Proc. of Interspeech, Portland, US, 2012.

[14] Charlet, D. and Barras, C. and Lienard, J.S., "Impact of Overlapping Speech Detection on Speaker Diarization for Broadcast News and Debates", in Proc. of ICASSP, 2013.

[15] Ghaemmagham, H. and Dean, D. and Sridharan, S., "Speaker Attribution of Australian Broadcast News Data", in Proc. of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), France, August 22-23, 2013.

[16] Larcher, A. and Bonastre, J.-F. and Fauve, B. and Lee, K. A., and Levy, C. and Li, H. and Mason, J. S. D. and Parfait, J.-Y., "ALIZE 3.0 - Open source toolkit for state-of-the-art speaker recognition", in Proc. of Interspeech, Lyon, France, 2013.