Semi-Supervised and Unsupervised Data Extraction Targeting Speakers: From Speaker Roles to Fame?

Carole Lailler, Grégor Dupuy, Mickael Rouvier, Sylvain Meignier

▶ **To cite this version:**

Carole Lailler, Grégor Dupuy, Mickael Rouvier, Sylvain Meignier. Semi-Supervised and Unsupervised Data Extraction Targeting Speakers: From Speaker Roles to Fame?. Interspeech satellite workshop on Speech, Language and Audio in Multimedia (SLAM), 2013, Marseille, France. hal-01433450

HAL Id: hal-01433450
https://hal.science/hal-01433450

Submitted on 1 Apr 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Semi-Supervised and Unsupervised Data Extraction Targeting Speakers: From Speaker Roles to Fame?

Carole Lailler, Grégor Dupuy, Mickael Rouvier, Sylvain Meignier

LUNAM Université, LIUM, Le Mans, France

first.lastname@lium.univ-lemans.fr

Abstract

Speaker identification is based on classification methods and acoustic models. Acoustic models are learned from audio data related to the speakers to be modeled. However, recording and annotating such data is time-consuming and labor-intensive. In this paper we propose to use data available on video-sharing websites like *YouTube* and *Dailymotion* to learn speaker-specific acoustic models. This process raises two questions: on the one hand, which are the speakers that can be identified through this kind of knowledge and, in the other hand, how to extract these data from such a noisy corpus that is the Web. Two approaches are considered in order to extract and to annotate the data: the first is semi-supervised and requires a human annotator to control the process, the second is totally unsupervised. Speakers models created from the proposed approaches were experimented on the REPERE 2012 TV shows test corpus. The identification results have been analyzed in terms of speaker roles and *fame*, which is a subjective concept introduced to estimate the ease to model speakers.

Index Terms: speaker identification, JFA, semi- and unsupervised speaker modeling, speaker roles, *fame*

1. Introduction

REPERE is a French evaluation campaign in the field of multimedia people in television documents. The main purpose of this challenge is to answer the questions "who is speaking ?" and "who is seen ?" at any time of the videos. The targets are both television professionals and guests, which can refer either to experts in a specific field, or to politicians, or celebrities. This paper is only concerned in the "who is speaking?" question. In this context, the identification task aims to determine the identity of the speakers, at any time.

The system presented in this paper is a two-levels architecture that uses both speaker diarization and speaker identification to process the shows. The speaker diarization level aims to partition the input audio stream into homogeneous segments, and group these segments according to the identity of the speakers. The purpose of the speaker identification level is to annotate the segments with the true identity of the speakers. However, available data in the training corpus are insufficient to learn specific and robust speaker models for each of the persons appearing in the videos: the coverage in the training corpus, in terms of number of speakers, is too low.

A solution to address the problem of insufficient coverage is to enhance the training corpus with data matching persons who are not already present. Nevertheless, the creation of such annotated corpora is time-consuming and labor-intensive. With the advent of video-sharing websites on the Internet, like *YouTube* and *Dailymotion*, it is now possible to collect innumerable data on speakers. The downside is that such data are noisy and poorly annotated, only the title and the description of the videos help to determine the topic: find satisfactory video content is not easy because either no information have been provided, or information are untrue, inaccurate or incomplete. In addition, the available videos are often of different qualities: some of them are professional videos while the others look like homemade movies shared by amateurs. Also, different recording situations (indoor or outdoor, with one person or with a group, ...) make data mining challenging: it is easier to exploit data from a politician show where a single man appears on the screen than data from a show where many speakers are interacting.

The use of Internet to build up corpora has lately been the subject of many research, especially in the field of speaker identification. In the video field, various works attempted to associate names to faces for a special type of web pictures. [1, 2, 3, 4, 5] focused on the face-name association in news photographs. [1] and [6] applied a face detector on the pictures and a named entity detector on the captions, then tried to find associations between detected names and faces. In the audio field, the main method focuses on learn a consistent association of speech and face from videos [7]. All the proposed approaches were focused on unsupervised methods applied to the identification of celebrities.

In this paper, two methods are proposed (semi-supervised and unsupervised) to build up specific speaker models from data from video-sharing websites and thus, to get round teh lack of data. The videos are retrieved using a list of speakers that may appear in the TV news. The semi-supervised approach needs a human annotator in order to control the automatic extraction of the data that are supposed match the targeted speaker. Thus, human interventions are greatly minimized. The unsupervised approach allows to automatically extract the data corresponding to the targeted speaker without any control from the human annotator. These methods are evaluated with the TV shows that compose the test corpus of the French evaluation campaign REPERE 2012. The evaluation focuses on the quality of the speaker models extracted from the data obtained through the semi- and unsupervised approaches. In addition, an analysis based on the subjective concept of people *fame* was conducted to understand relationship between speaker roles and identification results.

In the next section, we briefly describe the initial training and test corpora. Then, we present the semi- and unsupervised approaches used to model speakers using non-annotated data in section 3. The implementation of the two-levels architecture is described in section 4, and evaluation metrics as well as experiments results are given in section 5. In the section 6, the aim is to answer to the question: what is the nature of modeled people? by an analysis of the speaker role and *fame*. This section

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

is followed by some conclusions.

2. Corpus

This work in speaker identification was conducted as part of the REPERE 2012 evaluation campaign [8]. As such, experiments were performed on the test corpus of this evaluation campaign, composed of 3 hours of data. This data are drawn from 28 TV shows, recorded from French TV channels: BFM and LCP. The corpus is balanced between prepared speech, with 7 broadcast news from French radio stations, and spontaneous speech, with 21 political discussions or street interviews. Only a part of the recordings are annotated, giving a total duration of 3 hours.

The purpose of this work is to identify people who frequently appear in the news, so a list of 580 people was manually built with either people appearing in the media, or people likely to be present in the news, people who might appear. This list contains anchors, journalists, celebrities such as ministers, actors, singers, *etc.* 152 people from this list can be modeled using the training corpus. The training corpus is composed of every annotated data distributed during the French ESTER-2, ETAPE and REPERE evaluation campaigns. Among the 152 extracted models, 30.1% match people present in the test corpus.

Despite the amount of annotated data used as training corpus, 428 people from the list can not be modeled. External data is needed. Thus we propose to use data available on video-sharing websites like *YouTube* and *Dailymotion* to learn speaker-specific acoustic models.

3. Data extraction and speaker modeling

Video-sharing websites provides access to a considerable amount of data. However, data mining is challenging because of various factors: the quality of the media, the recording situation (indoor/outdoor, single speaker/group of speakers, *etc.*), the quality of annotations (inaccurate, incomplete or nonexistent). Building up a corpus is performed in two steps: data extraction then data annotation. The extraction is performed by retrieving videos on the video-sharing websites according to a request. The process is as follows:

1. *Request:* the request is composed of the name of the speaker to be modeled
2. *Filter:* all the videos in which the title do not include the name of the speaker to be modeled are put aside
3. *Download:* the first twenty videos are downloaded

Two different approaches are presented to annotate these data in terms of speaker identity, in order to learn speaker-specific acoustic models. The unsupervised method automatically takes the decisions. The semi-supervised method involves a human annotator to help the choices made by the system.

3.1. Unsupervised

The unsupervised approach aims for automatically select the segments that match the targeted speaker without any control from the human annotator. The main difficulty is to automatically detect if the person talking is the one sought. We made the assumption that the targeted speaker participate in each of the video extracted from the video-sharing websites. Indeed, this assumption has been validated on a portion of the training corpus listening to selected segments (the segments of less than 300 frames were not taken into account in this corpus).

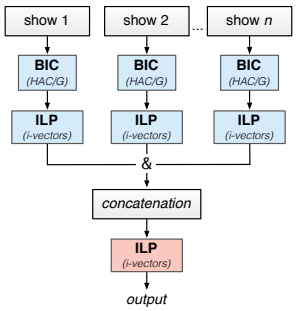

Figure 1: The audio cross-show speaker diarization architecture used to identify the cross-show speakers among the collection of videos.

Based on this assumption, an audio cross-show speaker diarization system is used to detect the speakers appearing across the multiple videos of the collection [9, 10, 11]. As illustrated in Figure 1, the cross-show diarization system first processes each video individually, by using a single-show speaker diarization system based on a Bayesian Information Criterion (BIC) segmentation followed by a clustering expressed as an Integer Linear Programming (ILP) problem. Then, the system attempts to identify speakers reappearing in several videos within the collection, by performing an overall ILP clustering [11].

After this cross-show diarization process, only the main cluster is considered. A filtering step is then performed to stop the process if not enough data are available to create the speaker model: if the speaker associated to the main cluster is appearing at least in three videos, and if the length of it interventions is longer than 2 minutes, then the acoustic model is created.

3.2. Semi-supervised

The aim of the semi-supervised method is to annotate the data extracted from the video-sharing websites while minimizing human efforts. We assume that the speaker to be modeled is present in each of the video collected, and that this speaker is the one who talk the most. An audio single-show speaker diarization system, as described in section 4.1, is run on each of the video. In order to correct the resulting clustering, that may not be perfect, a human annotator has to verify and invalidate the erroneous clusters. The purpose is to obtain the maximum number of segments that represent the targeted speaker, while maximizing the purity of the data by putting aside erroneous clusters. To minimize the human annotator effort, a validation of the audio segmentation according to the corresponding image is proposed: we have considered that the person appearing on the image is the one who is speaking because in the REPERE training corpus, the targeted speaker appears in about 80% of cases. The full process of the semi-supervised is as follow:

1. An audio speaker diarization system is run on each video,
2. Only the main cluster is considered (making the assumption that the main cluster match to the targeted speaker),
3. The images in the middle of each of the segment from the main cluster are extracted,
4. Human annotator verifies the speaker clustering by invalidating segments (so the picture) that do not contain the targeted speaker.

Figure 2: The application that help the annotator to invalidate audio segments according to the person appearing on the image.

An application has been developed to help the human annotator, by clicking on the images provided, to invalidate the segments that do not show the targeted speaker (Figure 2). More than 1900 hours of videos have been downloaded, and it took 224 hours to process the annotation. Finally, 480 hours of data have been annotated with the speaker identities. The ratio between the duration of the data to be annotated and the duration of the annotation itself is about 0.11. In [12], the manual annotation of a 2h08 corpus lasted 1h17, the ratio was about 0.60. Although it is less accurate, the duration of the annotation process, with the *semi-supervised* method is 6 times faster than a fully manual annotation.

4. Architecture of the identification system

In this paper we present a two-levels architecture that combines a speaker diarization system with a speaker identification system. The speaker diarization task aims to answer the question "who spoke, when ?", by partitioning an input audio stream into segments, and by clustering those segments according to the identity of the speakers. Experiments were carried out using the *LIUM_SpkDiarization* toolkit[1]. The speaker identification system consist in identifying each of the clusters with the real name of the speaker. This system is based on Joint Factor Analysis (JFA).

4.1. Speaker Diarization

The speaker diarization system is composed of an acoustic Bayesian Information Criterion (BIC) segmentation followed by a BIC hierarchical clustering using BIC both as similarity measure between speakers and as stop criterion for the merging process. Each speaker is modeled by a Gaussian distribution with a full covariance matrix. A Viterbi decoding is used to adjust the segment boundaries using Gaussian Mixture Models (GMMs) with 8 diagonal components, trained by Expectation-Maximization (EM) algorithm on the data of each speaker. Segmentation, clustering and decoding are performed using 12 MFCC+E, computed with a 10ms frame rate. Music and jingle regions are removed using Viterbi decoding with 8 one-state HMMs: 1 music model, 1 jingles model , 2 silence models (wide/narrow band), 1 narrow band speech model, and 3 wide band speech models (clean/over noise/over music). Each state is represented by a 64 diagonal GMM.

[1] http://www-lium.univ-lemans.fr/en/content/liumspkdiarization

In the previous steps, features were used unnormalized (to preserve information on the background environment). At this point, each speaker is not necessarily represented by a single cluster. The contribution of the background environment is removed through a feature normalization and the system then performs an ILP clustering dealing with i-vectors speaker models [13].

In order to identify the cross-show speakers in the *unsupervised* method, a final ILP clustering is performed from the concatenation of the single-show diarization outputs [11].

4.2. Speaker Identification

The speaker identification system aims to identify the real name of speaker for each cluster given by the speaker diarization system. The speaker identification system is based on the Joint Factor Analysis (JFA) framework [14, 15]. The purpose of JFA is to decompose the speaker-specific model into three different components: a speaker-session-independent component, a speaker-dependent component and a session-dependent component (each recording corresponding to one of these session). A supervector is defined as the concatenation of the GMM means components. Let D be the dimension of the feature space, the dimension of a supervector mean is $M.D$, where M is the number of components in the GMM. For a speaker s belonging in session h, the factor analysis model can be formulated as:

$$\mathbf{m}_{(h,s)} = \mathbf{m} + \mathbf{D}\mathbf{y}_s + \mathbf{U}\mathbf{x}_{(h,s)}, \quad (1)$$

where $\mathbf{m}_{(h,s)}$ is the session-speaker dependent supervector mean, $\mathbf{D}$ is $M.D \times M.D$ diagonal matrix, $\mathbf{y}_s$ the speaker vector (a $M.D$ vector), $\mathbf{U}$ is the session variability matrix of low rank R (a $M.D \times R$ matrix), and $\mathbf{x}_{(h,s)}$ are the channel factors, a R vector. All parameters of the JFA model are estimated by using the Maximum Likelihood criterion and the EM algorithm. Several sessions corresponding to each speaker have to be used for an accurate estimation of JFA parameters. 60-dimensional acoustic features were computed, with a 10ms frame rate. The features are composed of 19 MFCCs + log energy, and augmented by their first and second-order derivatives. The GMM-UBM is a gender- and channel-independent GMM composed of 1024 Gaussians. The dimension of R is 40.

5. Experiments

Experiments were performed on the test corpus of the REPERE 2012 evaluation campaign. This corpus is composed of 3 hours of data, drawn from 28 TV shows, recorded from French TV channels: BFM and LCP. The corpus is balanced between prepared speech, with 7 broadcast news from French radio stations, and spontaneous speech, with 21 political discussions or street interviews. Only a part of the recordings are annotated, giving a total duration of 3 hours.

5.1. Evaluation metrics

The Diarization Error Rate (DER) is the metric used to measure performance in the speaker diarization task. DER was introduced by the NIST as the fraction of speaking time which is not attributed to the correct speaker using the best match between references and hypothesis speaker labels.

The evaluation metric chosen to measure identification performance is the official REPERE Estimated Global Error Rate (EGER). This metric is defined as follow:

	Supervised	Supervised + Semi-supervised	Supervised + Unsupervised	Semi-supervised	Unsupervised
BFMStory	58.2%	**55.8%**	56.5%	90.3%	89.1%
CultureEtVous	56.1%	**53.7%**	**53.7%**	100.0%	100.0%
CaVousRegarde	62.4%	**56.4%**	**56.4%**	90.1%	90.1%
EntreLesLignes	**13.5%**	**13.5%**	**13.5%**	40.8%	63.5%
LCPInfo	52.7%	**50.5%**	51.1%	65.6%	89.1%
PileEtFace	51.2%	**22.3%**	42.1%	45.8%	66.1%
TopQuestions	35.3%	35.3%	**35.2%**	41.3%	53.2%
# of speaker models	152	410	397	377	343
% useful in the test corpus	30.1%	40.4%	39.7%	28.7%	23.9%
REPERE	46.5%	**41.9%**	44.2%	67.7%	77.2%

Table 1: EGER on the REPERE 2012 test corpus with the *semi-* and *unsupervised* methods, combined or not with the speaker models from the training corpus (*supervised* method). The number of speaker models extracted, as well as the coverage (% of speaker models really matching a speaker in the test corpus), are also presented.

$$EGER = \frac{\#fa + \#miss + \#conf}{\#total} \quad (2)$$

where $\#total$ is the number of person utterances to be detected, $\#conf$ the number of utterances wrongly identified, $\#miss$ the number of missed utterances and $\#fa$ the number of false alarms. Both DER and EGER are computed using the scoring tool developed by the LNE[2] as part of the ETAPE and the REPERE campaigns.

5.2. Speaker diarization results

Single-show Diarization Error Rates obtained on the REPERE 2012 test corpus are reported in Table 2. DER of each show was computed from the output of the first level of the architecture presented in Paragraph 4. The variability of the results directly depends on the type of video processed. The DER is approximately 7% on broadcast news videos (BFM story, LCP Info), 11% to 16% on political discussions videos, and 28% on people/entertainment videos (Culture Et Vous). This system obtained the best results during the ETAPE 2012 and REPERE 2013 evaluation campaigns [16].

	%Miss	%F.A.	%Sub.	%DER
BFMStory	0.48	1.45	5.91	7.86
CultureEtVous	4.21	2.99	21.74	28.95
CaVousRegarde	2.02	0.10	12.78	14.91
EntreLesLignes	0.00	0.46	11.23	11.70
LCPInfo	0.42	0.95	5.97	7.35
PileEtFace	0.04	0.39	16.27	16.71
TopQuestions	1.34	3.04	10.60	14.99
REPERE	0.95	1.41	9.92	12.30

Table 2: Single-show DER on the REPERE 2012 test corpus.

5.3. Speaker identification results

Estimated Global Error Rates obtained on the REPERE 2012 test corpus are presented in Table 1. The "supervised" column shows the results obtained with the 152 speaker models extracted from the training corpus. Other columns present results obtained with the *semi-* and *unsupervised* methods, combined or not with the speaker models from the training corpus. EGER of the *semi-* and *unsupervised* methods, when combined with speaker models from the *supervised* method, are 41.9% and 44.2%, respectively.

EGER obtained with both method is improved because of the increase of speaker models. The *supervised+semi-supervised* method gives the best results. The resulting speaker models are more robust because of the verification made by the human annotator. The *unsupervised* method (without the *supervised* data) gives a coverage of 23.9% (343 speaker models were automatically extracted), and a EGER of 77.2 %.

[2]The French National Laboratory of Metrology and Testing

6. Speaker roles influence

Two methods were proposed (*semi-supervised* and *unsupervised*) to increase the number of speaker models, or improve the existing models. This section present an analysis which focuses on the relationship between the speakers models and the role of the speakers.

6.1. Roles description

Five roles are described in the REPERE evaluation campaign which are commonly used in the literature [17, 18]. In this analysis, R4 and R5 have been merge because of their similarity.

- **R1**: The anchors. These speakers are characterized by their presence throughout the show, without discontinuity.
- **R2**: The journalists. They are TV professionals appearing one time or more during the show.
- **R3**: The reporters. Similar to the role R2, they are correspondents covering events outside the set of the show.
- **R4+R5**: The guests (R4). They are invited to interact with the actualities. They were asked for their knowledge or their *fame* to discuss under the guidance of the anchor. They are neither part of the organization committee, nor the leaders of debates. They can be present in different TV shows, especially during a highly publicized event. R5 role refers to everyone else that could appear, like interviewed people in a report.

6.2. Results and comments

Table 3 shows the EGER and the coverage (% of speaker models really matching a speaker in the test corpus) of each of the roles (R1, R2, R3 and R4+R5), for each of the systems that have been presented in paragraph 5.3. The column "Reference" only shows the role distribution of the manually built list of 580 speakers used to collect the videos from the video-sharing websites. For example, this list contains 90.9% of anchors (i.e. R1 role) who are present in the test corpus.

Regarding the *supervised* system, a EGER of 8.6% and 12.2% were obtained for the R1 and the R2 roles, respectively. These low error rates are essentially due to the presence of the R1 and R2 speakers both in the training and test corpora of

	Reference	Supervised	Supervised + Semi-supervised	Supervised + Unsupervised	Semi-supervised	Unsupervised
R1	*(90.9%)*	**8.6%** *(81.8%)*	9.5% *(81.8%)*	9.0% (81.8%)	83.8% *(18.1%)*	100.0% *(0.0%)*
R2	*(85.7%)*	**12.2%** *(85.7%)*	**12.2%** *(85.7%)*	**12.2%** *(85.7%)*	43.2% *(42.8%)*	65.6% *(28.5%)*
R3	*(50.0%)*	**43.4%** *(50.0%)*	**43.4%** *(50.0%)*	**43.4%** *(50.0%)*	100.0% *(0.0%)*	100.0% *(0.0%)*
R4+R5	*(35.9%)*	64.7% *(20.1%)*	**57.5%** *(32.4%)*	61.1% *(30.2%)*	63.0% *(31.5%)*	70.7% *(28.0%)*
# of speaker models	512	152	410	397	377	343

Table 3: EGER comparison between the roles (R1, R2, R3 and R4+R5) and the speaker models of each system in the REPERE 2012 test corpus. Values in parentheses indicate the number of speakers with the corresponding role divided by the number of speaker models in each system.

the REPERE campaign. R3 and R4+R5 EGER are 43.4% and 64.7%, respectively. These rates are consistent with the frequency of appearance of the corresponding speakers. Less the speaker takes part in the training corpus, more it is difficult to detect him in the test corpus. 81.8% of R1 speaker of the test corpus have a model (85.7% for R2 speaker). It is particularly true for the R4+R5 speakers, corresponding to the guests in a broad sense, 20.1% of those speakers have a model.

Supervised+semi-supervised and *supervised+unsupervised* methods allow to better detect the R4+R5 role. Compared to the *supervised* method, the EGER of the *supervised+semi-supervised* decrease from 64.7% to 57.5% (-7.2% absolute), and the EGER of the *supervised+unsupervised* decrease from 64.7% to 61.1% (-3.6% absolute). The difference between the two methods is explained by the fact that the *Supervised+semi-supervised* method have more data to learn models. Indeed, the models coming from the *Supervised* method is learned with more data. Moreover, 14 new speakers models are added.

We introduce the subjective notion of *fame* of a speaker. A speaker has a significant *fame* if his presence on TV is going to beyond the scope of a channel. The celebrities, politicians or artists are easily recognizable by their large representation in various shows: their interviews are widely diffused. Thus, they have a wide *fame*. On the other hand, people like TV professionals, only appearing in the TV channel they work for, have a limited *fame*. It is easy to find data on video-sharing websites for famous people; it is difficult for not famous ones.

Thus, in the list of 580 speakers we have build, 9.1% [3] of R1 role misses in order to obtain all the anchors, 14.3% of R2 role for the journalists, 50% of R3 role for reporters and 64.1% of R4+R5 roles for the guests. Experiments show that the *unsupervised* method does not help to identify a single anchor, and the *semi-supervised* method has only found 18.1% of them. The half of the speakers list of 580 speakers is labelled as R3 role, and none of the two methods helps to identify this category of people (0%). Conversely, R4+54 roles (guests) only represents 35.9% of the 580 speakers list. The *semi-supervised* method allows to successfully identify 31.5% of the guests, and the *unsupervised* method, in which speakers models are automatically created without any human supervision, is able to identify 28% of them.

Aside from the headliners, the broadcast news programs are uniquely composed of TV professionals who only officiate on that channel. The influence of these individuals is lower and it becomes more difficult to trace, because their name is often associated with a single channel or to a single show. They always appear in the same situation, and fulfill the same role each time. This set includes the R1, R2 and R3 roles.

However, differences can be identified within these three following roles:

- **R1**: They always appear with the same appearance and the same clothes, in the same context. They finally have a relative important *fame*: again, it is difficult to obtain relevant and reliable data to produce robust speaker models with the proposed method.
- **R2**: This role is ultimately less difficult to identify because journalists often have a presence on several channels, in different situations. This variability leads to a better identification. The fact that they appear on several channels increase their *fame*. For example, the Semi-supervised method has a recovery rate of 42.8%.
- **R3**: This role corresponds to individuals usually appearing outside, and very occasionally. That imply real difficulties to obtain relevant data. In addition, little information flows because the contexts in which they appear are usually very different. Moreover, these individuals often work in noisy environments which increase the difficulties to obtain reliable acoustic information.

[3] This percentage comes from table 3, it corresponds to 100%-90.9%, etc.

7. Conclusions

In this paper, the various approaches proposed help to quickly obtain data in order to produce models for speaker identification. Regardless the method used, people with an important *fame* like celebrities are easy to model because of the ease to find related data. However, the proposed methods do not provide sufficient data to model anchors or journalists (unless they have an activity outside the channel). The *semi-supervised* approach has obtained better results than the *unsupervised* approach. Speaker models produced are more robust because of the controls made by the human annotator.

Aside from celebrities which are true "headliners", the shows are composed of a group of TV professionals who officiate on the channel in question (sometimes exclusively). The *fame* of these persons is limited; it becomes difficult to find related data on video-sharing websites because their presence is unfrequent, and restricted to a particular situation.

8. References

[1] T. L. Berg, A. C. Berg, J. Edwards, and D. A. Forsyth, "Whos in the picture," in *NIPS*, 2004.

[2] D. Ozkan and P. Duygulu, "A graph based approach for naming faces in news photos," in *CVPR (2)*, 2006, pp. 1477–1482.

[3] M. Guillaumin, T. Mensink, J. J. Verbeek, and C. Schmid, "Automatic face naming with caption-based supervision," in *CVPR*, 2008.

[4] P. T. Pham, M.-F. Moens, and T. Tuytelaars, "Cross-media alignment of names and faces," *IEEE Transactions on Multimedia*, vol. 12, no. 1, 2010.

[5] J. Luo, B. Caputo, and V. Ferrari, "Who's doing what: Joint modeling of names and verbs for simultaneous face and pose annotation," in *NIPS*, 2009, pp. 1168–1176.

[6] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y. W. Teh, E. G. Learned-Miller, and D. A. Forsyth, "Names and faces in the news," in *CVPR (2)*, 2004, pp. 848–854.

[7] M. E. Sargin, H. Aradhye, P. J. Moreno, and M. Zhao, "Audiovisual celebrity recognition in unconstrained web videos," in *ICASSP*, 2009, pp. 1977–1980.

[8] J. Kahn, O. Galibert, M. Carré, A. Giraudel, P. Joly, and L. Quintard, "The repere challenge: Finding people in a multimodal context," in *Odyssey 2012 - The Speaker and Language Recognition Workshop*, 2012.

[9] V.-A. Tran, V. B. Le, C. Barras, and L. Lamel, "Comparing multistage approaches for cross-show speaker diarization," in *Proceedings of Interspeech*, Florence, Italy, 2011.

[10] Q. Yang, Q. Jin, and T. Schultz, "Investigation of cross-show speaker diarization," in *Proceedings of Interspeech*, Florence, Italy, 2011.

[11] G. Dupuy, M. Rouvier, S. Meignier, and Y. Estève, "I-vectors and ILP clustering adapted to cross-show speaker diarization," in *Proceedings of Interspeech*, Portland, Oregon (USA), 2012.

[12] Y. E. Thierry Bazillon and D. Luzzati, "Manual vs assisted transcription of prepared and spontaneous speech," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), may 2008.

[13] M. Rouvier and S. Meignier, "A global optimization framework for speaker diarization," in *Odyssey Workshop*, Singapore, 2012.

[14] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 345–354, 2005.

[15] D. Matrouf, N. Scheffer, B. Fauve, and J.-F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification," in *Proc. Interspeech*, 2007.

[16] O. Galibert and J. Kahn, "The first official repere evaluation," in *SLAM 2013*, France, Marseille, 2013.

[17] R. Barzilay, M. Collins, J. Hirschberg, and S. Wittaker, "The rules behind roles: Identifying speaker role in radio broadcasts," in *Proceedings of the National Conference on Artificial Intelligence*, 2000, pp. 679–684.

[18] T. Bazillon, B. Maza, M. Rouvier, F. Bechet, and A. Nasr, "Speaker role recognition using question detection and characterization," in *Interspeech*, 2011.