# **Speaker Role Recognition on TV Broadcast Documents**

Benjamin Bigot, Corinne Fredouille

University of Avignon, LIA, France

firstname.lastname@univ-avignon.fr

### Abstract

In this paper, we present the results obtained by a state-ofthe-art system for Speaker Role Recognition (SRR) on the TV broadcast documents issued from the REPERE Multimedia Challenge. This SRR system is based on the assumption that cues about speaker roles may be extracted from a set of 36 low level features issued from the outputs of a Speaker Diarization process. Starting from manually annotated speaker segments, we first evaluate the performance of the SRR system, formerly evaluated on Broadcast radio recordings, on this heterogeneous set of TV shows. Consequently, we propose a new classification strategy, by observing how building show-dependent models improves SRR. The system is then applied on some speaker segmentation outputs issued from an automatic system, enabling us to investigate the influence of the errors introduced by this front-end process on Role Recognition. In these different contexts, the system is able to correctly classify 86.9% of speaker roles while being applied on manual speaker segmentations and 74.5% on automatic Speaker Diarization outputs.

**Index Terms**: speaker role recognition, speech processing, content-based indexing of audiovisual documents.

#### 1. Introduction

Maintaining efficient means of access to the information held in the huge mass of audiovisual documents broadcast everyday by TV and radio channels is very challenging. The increasing number of projects and evaluation campaigns puts to the fore the important work currently achieved in order to propose automatic methods dedicated to information extraction, contentbased indexing and structuring in audiovisual documents.

The REPERE Multimedia Challenge [1] (2010-2014) is dedicated to the specific task of person identification in TV programs. It provides a framework (corpora and evaluation protocols) to support research on this topic in multimodal conditions. The work presented in this paper is part of the PERCOL project which is one of the three consortia chosen to participate at this challenge. In the context of the first official phase of the REPERE campaign, the scientific partners involved in PER-COL have proposed several systems dedicated to speaker identities through the recognition of pronounced names and speaker identification in speech, person name detection in overlaid texts and face recognition in video.

We are interested in bringing information relative to speaker roles in a speaker identification perspective. In well structured documents, as in broadcast news programs, several studies have already taken advantage of the links between speaker roles (like anchor, journalist, guest or interview participant) and content structure. Role information has previously been used in [2] to summarize broadcast news documents, for topic indexing [3] and for story segmentation, relying either on the detection of the anchorman [4] or journalists [5]. Delphine Charlet

Orange Labs, Lannion, France

delphine.charlet@orange.com

This paper is dedicated to the application of a state-of-theart speaker role recognition system on the document set of the REPERE challenge since there are several potential benefits in using information brought by speaker roles in a speaker identification perspective. For example, a TV show presenter is expected to introduce his guests and chroniclers by citing their names. Speaker roles may as well bring confidence in the outputs of a speaker recognition process.

The paper is structured as follows. In section 2 we present a brief description of related works on speaker role recognition. The document set as well as both the Speaker Diarization and Speaker Role Recognition systems used in this study are presented in sections 3 and 4 respectively. Section 5 is dedicated to the experiments carried out. Finally, we conclude this paper with some perspectives.

# 2. Related Works

First contributions to SRR [6, 7] are methods based on the outputs of Automatic Speech Recognition (ASR). A second category of approaches concern works based on Social Network Analysis and Social Affiliation Network applied on Speaker Diarization [8, 9]. In this second case, prior knowledge about the structure of the show is taken into account to determine relevant roles. These methods are mainly based on three classic roles: anchorman, journalist, other (or guest). The number of roles could be greater, but on very specific corpora (bulletins from a same news program). A more detailed survey of the state-of-the-art concerning these methods can be found in [10]. More recent contributions tend to benefit from speech transcriptions as well as from the temporal organization of speaker turns. In [11], authors integrate information relative to speaking style and a priori information about turn-taking patterns of conversations in a Dynamic Bayesian Network (DBN). The method in [12] assumes that cues about speaker roles are available in the way speakers formulate their questions. In [13] both structural and lexical features are used together. The approach proposed in [14] classifies speaker segments among the three classical roles. A first step based on the temporal distribution of speech segments and on the average Bayesian Information Criterion realizes the detection of anchorman. A second step, based on textual information achieves the classification of journalist and other. In [15] the authors investigate the links between speech spontaneity markers and speaker roles. The work of [16] is based on prosodic and temporal features calculated for every speaker segment obtained from a Speaker Diarization. The decision step is achieved using Conditional Random Fields (CRF). Because of the important diversity among these contributions (methods, features, language, corpus and metrics), these results are difficult to compare together.

# 3. Corpus

At each step of the REPERE challenge, a set of TV broadcast documents and its annotations is delivered to the participants. The training set of the first phase has been annotated with speaker roles. We present in details this corpus as well as its features from a speaker role recognition point of view.

### 3.1. REPERE Broadcast TV document set

This data set is composed of 135 documents corresponding to several recordings of 7 different TV programs taken from two different TV channels. It corresponds to an overall speech duration of 24 hours distributed among these different programs as reported in Table 1. This set can be divided into three categories of programs :

- **Broadcast News** (13.8h) with *BFMStory*, *Showbiz* and *LCPInfo*. Documents belonging to this category count for more than the half of the entire data set duration. The program *Showbiz* is slightly different from the other ones and is a People News and gossip program.
- **Debates** (6.6h) : among this set, *EntreLesLignes* is dedicated to journalistic questions, while *ÇaVousRegarde* focuses on society questions. *PileEtFace* is a head-to-head political debate.
- The last category (3.6h) is for *TopQuestions*. These documents are recordings of the parliamentary sessions of the French National Assembly.

| Name           | nb. of<br>shows | doc.<br>type | speech<br>duration |
|----------------|-----------------|--------------|--------------------|
| BFMStory       | 14              | News         | 7.9h               |
| Showbiz        | 66              | News         | 1.9h               |
| LCPInfo        | 15              | News         | 4h                 |
| ÇaVousRegarde  | 6               | Debate       | 2.2h               |
| EntreLesLignes | 7               | Debate       | 2.2h               |
| PileEtFace     | 9               | Debate       | 2.2h               |
| TopQuestion    | 18              | Nat. Ass.    | 3.6h               |
|                | 135             |              | 24h                |

Table 1: The REPERE data set and its speech distribution among various programs

#### 3.2. Role definitions

Manual speaker segmentation has been enriched with annotations relative to speaker roles. Five types of roles have been consensually defined by the REPERE participants:

- **type R1** is for the anchorman persons, presenters and TV newscasters. Only one R1-type person is typically expected in each program, except for *Showbiz* where the shows are presented by 2 anchormen. As presented in table 2, R1 is the only role present in all the programs.
- **type R2** is for journalists and chroniclers. According to the chosen definition, these speakers must appear physically on the television studio set. We can see in tables 2 and 3 that these speakers are present in only three types of programs.

|                 | R1   | R2  | R3   | R4   | R5   | spk # |
|-----------------|------|-----|------|------|------|-------|
| BFMStory        | 6.2  | 4.8 | 16.5 | 19.4 | 53.1 | 273   |
| Showbiz         | 21.7 | 0   | 12.6 | 0    | 65.7 | 563   |
| LCPInfo         | 5.5  | 2.2 | 19   | 5.8  | 67.5 | 274   |
| ÇaVousRegarde   | 11.5 | 0   | 7.7  | 50   | 30.8 | 52    |
| EntreLesLignes  | 20   | 80  | 0    | 0    | 0    | 35    |
| PileEtFace      | 34.6 | 0   | 0    | 65.4 | 0    | 26    |
| TopQuestions    | 12.9 | 0   | 0    | 0    | 87.1 | 140   |
| overall (%)     | 14.2 | 3.5 | 12.6 | 8.2  | 61.5 |       |
| overall (# spk) | 194  | 47  | 172  | 112  | 838  | 1363  |

Table 2: speaker role proportion for every type of program and total number of speakers

|                | R1   | R2   | R3   | R4   | R5   |
|----------------|------|------|------|------|------|
| BFMStory       | 25.1 | 14.2 | 11   | 38.7 | 11   |
| Showbiz        | 14   | 0    | 52.1 | 0    | 33.9 |
| LCPInfo        | 27.2 | 3.8  | 20   | 27.8 | 21.2 |
| ÇaVousRegarde  | 29   | 0    | 3.3  | 63   | 4.7  |
| EntreLesLignes | 25.8 | 74.2 | 0    | 0    | 0    |
| PileEtFace     | 21.8 | 0    | 0    | 78.2 | 0    |
| TopQuestions   | 3.9  | 0    | 0    | 0    | 96.1 |
| overall (%)    | 21.6 | 12.1 | 11.3 | 30.4 | 24.6 |
| oveall (h)     | 5.17 | 2.91 | 2.72 | 7.28 | 5.90 |

Table 3: speech proportion in percent of the speech duration per program depending on speaker roles

- **type R3** is for journalists who are not present on the television studio set and for voice-over journalists during reports. This role occurs in the News programs and in the reports of *ÇaVousRegarde* (cf. Table 2).
- **type R4** stands in a large manner for the guests and the experts present in the shows, and more precisely to any person that is not working for the TV show. These speakers may be on the television studio set, or on live by telephone. As shown in Table 3, this is the most important role in terms of speech duration.
- type R5 contains all remaining speakers. It gathers anonymous persons, sound-bites and archives taken from press conference, spectators asking questions during a show, person interviewed during a report and politicians talking during a session at the National Assembly. This large category gathers 838 speakers but does not stand for the larger proportion of speech duration.

Speaker role distribution in terms of speaker numbers and speech duration is significantly different over the programs of the data set as depicted in Tables 2 and 3. For instance *BFM-Story* and *LCPInfo* are the only programs that contain five roles. On the contrary *EntreLesLignes*, *PileEtFace* and *TopQuestions* contain only two roles with an important disproportion in terms of speaker numbers and speech duration.

# 4. Speaker Role Recognition Architecture

The Speaker Role Recognition system used in this study has been previously presented in [10]. This systems has reached the good score of 92% of roles correctly attributed to the speakers of the broadcast radio programs composing the EPAC project corpus. It is initially dedicated to the recognition of 5 roles (*an-chorman, punctual and recurrent journalists*, and *punctual and recurrent others*). The terms *punctual* and *recurrent* characterize speakers activity in one given document. It has been adapted for the need of this study to the role categories presented above. We first briefly present the speaker diarization system used in this work and then depict the classification procedure applied to the SRR system.

### 4.1. Speaker Diarization

The diarization system used in this work is the one presented in [17]. 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.

# 4.2. Speaker Role Recognition system

This system relies on the assumption that cues about speaker roles can be extracted from a set of 36 low-level features (14 temporal features, 10 acoustic and 12 prosodic ones) computed from speech signal and from the temporal organization of speech turns available from a speaker diarization process. These features are used in a second time to model speaker roles using a supervised classification approach.

In [10] we put to the fore the efficiency of a hierarchical classification process where each classification step is reduced to a two-class problem. At each step, the redundancy or correlation of features for a given problem is reduced using a Principal Component Analysis and a discriminant model is learnt using a Linear SVM classifier. In this current study, the successive steps of the classification are adapted to the current problem. As presented in figure 1 a first step concerns the classification of the role R1. Then the classified speakers found as "not R1" are directed to a second classification step that considers R2 and R3 roles versus R4 and R5. Finally the last classification steps are done in parallel and separate R2 from R3 and R4 from R5.

### 5. Experiments

This classification is achieved at the scale of a speaker cluster and we assume that one speaker in one document has exactly one role. In order to deal with the quite limited number of samples in the corpus, the classification process is done in a leaveone-out fashion. Therefore, one document is used for test while the other documents are used to learn models.

|      |      | predicted |           |  |
|------|------|-----------|-----------|--|
|      |      | C1        | $\neg C1$ |  |
| ref  | C1   | TP        | FN        |  |
| ICI. | ¬ C1 | FP        | TN        |  |

Table 4: confusion matrix

Performance is reported in terms of Correct Classification Rate CCR defined according to the confusion matrix in Table 4:

$$CCR = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$



Figure 1: the speaker role recognition architecture

Recall and Precision measures defined respectively as follows are also reported :

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

in which TP, FP, and FN stand for true positive, false positive and false negative respectively.

### 5.1. Baseline-SRR using show-independent role models

This first experiment is done with manual speaker and role annotations, by applying a leave-one-out over the 135 documents of the data set independently of the type of program. Thus, while one document is processed for recognition, the 134 other documents are used for training.

The overall Correct Classification Rate (CCR) (cf. Table 5) is equal to 73.14% of speaker roles and 66.1% of the processed speech duration. A maximum 81.4% CCR is reached on the documents belonging to the program TopQuestions. The worst CCR value has been obtained on the programs Entre-LesLignes. Globally, best performance has been reached on broadcast News programs. Role recognition seems less efficient on the debate programs where precision and recall values are particularly low for the role R1. In the one hand, we can assume that debate presenters and News broadcasters do not share similar temporal, prosodic and acoustic characteristics. On the other hand, R1 speakers in news programs are more numerous than R1 in debates. This may have led to R1 models more adapted for news programs. We have also observed that all R1 speakers in Showbiz have been attributed to the role R5. This confusion may be caused by overlapping music during speech interventions of R1 speakers in Showbiz which makes them more similar to R5 speakers. Finally this baseline system presents several major issues. First, this system allows confusion between role types that do not exist in test document. For instance, while processing recognition on a TopQuestions document, recognition may conduct to attribute speakers to the roles R2, R3 or

|                |       | R     | 1    | R     | 2    | R     | 3    | R     | 4    | R     | 5    |
|----------------|-------|-------|------|-------|------|-------|------|-------|------|-------|------|
|                | CCR   | Prec. | Rec. |
| BFMStory       | 78.4  | 71.4  | 58.8 | 54.5  | 46.1 | 65.6  | 46.6 | 81.2  | 73.6 | 82.1  | 95.1 |
| Showbiz        | 73.7  | 0     | 0    | 0     | 0    | 92.3  | 67.6 | 0     | 0    | 71.8  | 99.2 |
| LCPInfo        | 77    | 80    | 33.3 | 0     | 0    | 71.4  | 19.2 | 60.9  | 87.5 | 79.5  | 98.9 |
| ÇaVousRegarde  | 48.1  | 30    | 50   | 0     | 0    | 40    | 50   | 75    | 23   | 50    | 87.5 |
| EntreLesLignes | 25.7  | 28.5  | 57.1 | 100   | 17.8 | 0     | 0    | 0     | 0    | 0     | 0    |
| PileEtFace     | 34.6  | 60    | 33.3 | 0     | 0    | 0     | 0    | 85.7  | 41.1 | 0     | 0    |
| TopQuestions   | 81.4  | 100   | 66.7 | 0     | 0    | 0     | 0    | 0     | 0    | 94.4  | 83.6 |
| overall (%)    | 73.14 | 60    | 18.6 | 52.4  | 23.4 | 62.8  | 47   | 69.9  | 58   | 75.8  | 95.9 |

Table 5: SRR performance in terms of Correct Classification Rate, Precision and Recall for every role and program type using the baseline architecture on manual speaker and role references

|                |      | R1    |      | R2    |      | R3    |      | R4    |       | R5    |      |
|----------------|------|-------|------|-------|------|-------|------|-------|-------|-------|------|
|                | CCR  | Prec. | Rec. | Prec. | Rec. | Prec. | Rec. | Prec. | Rec.  | Prec. | Rec. |
| BFMStory       | 83.5 | 100   | 58.8 | 80    | 61.5 | 76.5  | 57.8 | 80.8  | 79.2  | 85    | 97.9 |
| Showbiz        | 90.4 | 81.6  | 76.2 | X     | X    | 97.1  | 94.4 | Х     | Х     | 91.8  | 94.3 |
| LCPInfo        | 87.2 | 92.3  | 80   | 0     | 0    | 76    | 73.1 | 91.7  | 68.8  | 89.9  | 96.2 |
| ÇaVousRegarde  | 48.1 | 0     | 0    | X     | X    | 0     | 0    | 58.1  | 69.2  | 33.3  | 43.8 |
| EntreLesLignes | 85.7 | 66.7  | 57.1 | 89.7  | 92.9 | Х     | X    | X     | Х     | Х     | X    |
| PileEtFace     | 65.4 | 50    | 44.4 | X     | X    | Х     | X    | 72.2  | 76.64 | Х     | X    |
| TopQuestions   | 97.8 | 94.1  | 88.9 | Х     | X    | Х     | X    | Х     | Х     | 98.4  | 99.2 |
| overall (%)    | 86.9 | 82.7  | 71.6 | 85    | 72.3 | 85.6  | 76.2 | 74.3  | 75    | 89.7  | 95.1 |

Table 6: SRR performance in terms of Correct Classification Rate, Precision and Recall for every role and program type using a program dependent architecture on manual speaker and role references

R4 even if *TopQuestions* does not contain these roles. A second issue lays in the difference observed among speakers belonging to a given role type. For instance, we have observed low performance in R1 role recognition. There also exists an important confusion between R2 and R4 because these speakers seem to share similar characteristics. Finally, the size and the variety of speakers held in the R5 role may tend to unbalance the classification process. To overcome these limitations, we have first used program names as complementary features in the classification methods. Better performances have been reached in a second experiment, presented above, where role models are learnt depending on the type of programs.

#### 5.2. SRR using show-dependent role models

In this experiment, the leave-one-out process is limited to the documents corresponding to one given program. For instance, to recognize speaker roles in *TopQuestions* we only use the 18 documents available. This configuration reduces the amount of data used to model speaker roles. Compared to the previous experiment, one benefit provided by this program-dependent approach is to make impossible several role confusions since the classifier will only learn SVM models for roles that really occur in these programs.

Overall CCR reaches 86.9% of speakers as reported in Table 6. This corresponds to 83% of the duration correctly labelled. Performance has been globally improved for every show except for *ÇaVousRegarde*. In the latter, CCR value remains unchanged compared to the one reached in the previous experiment and all R1 and R3 speakers have been attributed to the roles R4 and R5. Globally we observe that in debate programs, the confusion between R1 speakers and the other roles remains important. This puts to the fore a possible lack of efficiency of the actual low level features used for these documents. Another explanation may stand in the fact that the speaker and role annotations provided for these programs do not correspond to entire shows. In the next experiment, we apply this program-dependent strategy to automatic speaker segmentations provided by the speaker diarization system described in section 4.

#### 5.3. SRR using show-dependent role models on Speaker Diarization outputs

Performance of the automatic Speaker Diarization system is reported in terms of Diarization Error Rate (DER). The overall performance on the corpus is equal to 12.1%, including overlapped speech in the evaluation. In Table 8 we observe that DER values depend on the type of programs. *Showbiz* presents the most important error rate value. This is mainly due to the high level of background music and noise, which gives a high miss detection rate, and also makes the clustering process more difficult for the detected speech. *EntreLesLignes* also presents an important DER value. This program contains several sequences of overlapping speech between chroniclers of the show.

To produce a ground truth for the evaluation of speaker role on the automatic speaker clusters, we first produce an alignment between the manual speaker segments and the outputs of the automatic speaker diarization system. Using the toolbox provided by NIST during a previous evaluation campaign we apply the Hungarian algorithm in order to associate automatic clusters with reference speakers. Then using these associations, we project the manual annotations for speaker roles over the automatic speaker clusters.

|                |      | R     | 1    | R     | 2    | R     | 3    | R     | 4    | R     | 5    |
|----------------|------|-------|------|-------|------|-------|------|-------|------|-------|------|
|                | CCR  | Prec. | Rec. |
| BFMStory       | 69.3 | 54.5  | 70.6 | 62.5  | 38.5 | 50.9  | 67.4 | 77.4  | 46.1 | 80.4  | 86   |
| Showbiz        | 80.1 | 61.1  | 81.5 | Х     | X    | 92.1  | 87.9 | X     | X    | 89.4  | 72.8 |
| LCPInfo        | 76.5 | 52.2  | 80   | 0     | 0    | 55.4  | 76.6 | 80    | 25   | 92.1  | 84.8 |
| ÇaVousRegarde  | 44.2 | 0     | 0    | Х     | X    | 0     | 0    | 51.6  | 60.9 | 31.3  | 50   |
| EntreLesLignes | 79.4 | 50    | 42.9 | 85.7  | 88.9 | Х     | Х    | X     | X    | Х     | Х    |
| PileEtFace     | 52   | 28.6  | 22.2 | Х     | X    | Х     | Х    | 61.1  | 68.8 | Х     | X    |
| TopQuestions   | 83.8 | 83.8  | 41.7 | 55.6  | X    | Х     | Х    | Х     | X    | 92.9  | 88.1 |
| overall (%)    | 74.5 | 53.9  | 65.9 | 78.4  | 64.4 | 66.5  | 76.9 | 65.4  | 49.5 | 86.7  | 83   |

Table 7: SRR performance in terms of Correct Classification Rate, Precision and Recall for every role and program type using a program dependent architecture on automatic speaker diarization outputs

| sources        | DER  |
|----------------|------|
| BFMStory       | 12.6 |
| Showbiz        | 34.4 |
| LCPInfo        | 8.9  |
| ÇaVousRegarde  | 12.2 |
| EntreLesLignes | 17.0 |
| PileEtFace     | 10.5 |
| TopQuestions   | 4.0  |
| total          | 12.1 |

Table 8: Diarization Error Rate for each program type

One consequence of the automatic process is that several speakers of the reference do not match an "automatic" speaker. These differences are reported in Table 9 and 10. We can see that among the 1363 speakers, only 885 are associated with a corresponding cluster. This loss is directly related to the DER since most of these lost speakers are from the *Showbiz* program. Considering the speech duration lost during this process, we have evaluated (cf Table 10) that it represents only 6.25% (1.5*h*) of the initial document set. The impact of the process is as well important on the speakers belonging to the class R1. Their overall speech duration in the outputs of the automatic clustering is equal to 4.29h instead of 5.17h in the reference data.

|                  | R1   | R2   | R3   | R4   | R5   | # spk |
|------------------|------|------|------|------|------|-------|
| BFMStory         | 7.6  | 5.8  | 19.1 | 23.1 | 44.4 | 225   |
| Showbiz          | 26.9 | 0    | 32.8 | 0    | 40.3 | 201   |
| LCPInfo          | 6.8  | 2.3  | 21.3 | 7.2  | 62.4 | 221   |
| ÇaVousRegarde    | 14   | 0    | 9.3  | 53.5 | 23.2 | 43    |
| EntreLesLignes   | 20.6 | 79.4 | 0    | 0    | 0    | 34    |
| PileEtFace       | 36   | 0    | 0    | 64   | 0    | 25    |
| TopQuestions     | 13.2 | 0    | 0    | 0    | 86.8 | 136   |
| overall (%)      | 14.2 | 5.1  | 18.1 | 12.1 | 50.5 |       |
| overall (number) | 126  | 45   | 160  | 107  | 447  | 885   |

Table 9: Speaker role in the diarization outputs in terms of speaker population

Speaker Role Recognition is then performed and an overall CCR equal to 74.5% is reached (cf Table 7). In term of speech duratio, it corresponds to 89% of the processed speech (22.5h) and 84.5% of the entire data set (24h). Program ranking in

|                | R1   | R2   | R3   | R4   | R5   | dur. |
|----------------|------|------|------|------|------|------|
| BFMStory       | 20.1 | 15   | 12.2 | 41   | 11.7 | 7.27 |
| Showbiz        | 14.6 | 0    | 66.2 | 0    | 19.2 | 1.51 |
| LCPInfo        | 24.9 | 3.7  | 19.9 | 29.7 | 21.8 | 3.77 |
| ÇaVousRegarde  | 27.9 | 0    | 3.4  | 64.9 | 3.8  | 2.08 |
| EntreLesLignes | 26.6 | 73.4 | 0    | 0    | 0    | 2.14 |
| PileEtFace     | 19   | 0    | 0    | 81   | 0    | 2.1  |
| TopQuestions   | 3.3  | 0    | 0    | 0    | 96.7 | 3.62 |
| overall (%)    | 19.1 | 12.4 | 12   | 31.8 | 24.6 |      |
| overall (h)    | 4.29 | 2.8  | 2.71 | 7.15 | 5.54 | 22.5 |

Table 10: Speaker role in the diarization outputs in terms of speech duration

terms of CCR is the same as the one observed on the manual segmentations. The best classification rate has been obtained on *TopQuestions* with 83.8% of speaker roles correctly attributed.

### 6. Conclusion

In this paper we have presented an experimental study dedicated to the application of a state-of-the-art Speaker Role Recognition system on the audiovisual document set provided in the context of the Multimedia Challenge REPERE. We first have carried out a baseline experiment by modelling speaker roles independently from the TV program processed. This system has been able to correctly attribute 73.14% of speaker roles, by using manual speaker segmentation. We have then proposed to build program dependent models of speaker roles. This second system has reported a correct classification rate of 86.9% on the same conditions. A third experiment has consisted in applying program-dependent models on the outputs of a speaker diarization system with a DER equal to 12.1%. On this difficult document set, our system has been able to correctly recognize 74.5%of speaker roles. In the context of the PERCOL project involved in the REPERE Challenge, our future work will be directed on the use of speaker role information in combination with speaker recognition system. We will as well investigate relations existing between speaker roles and the presence of person names pronounced in speech data in combination with systems like those presented in [18].

# 7. Acknowledgement

The authors thank the financial supports ANR 2010-CORD-102-02 from the French National Research Agency (ANR).

### 8. References

- Kahn, J., Galibert, O., Quintard, L, Carre, M., Giraudel, A. and Joly, P. "A presentation of the repere challenge," in CBMI, 2012, pp. 1–6.
- [2] Maskey, S. and Hirschberg, J. "Automatic speech summarization of broadcast news using structural features". in Proc. of EU-ROSPEECH 2003.
- [3] Amaral, R. and Trancoso, I. "Topic Indexing of TV Broadcast News Programs" in Proc. of PROPOR 2003.
- [4] Kolluru, B. and Gotoh, Y. "Speaker role based structural classification of broadcast news stories" in Proc. of INTERSPEECH 2007
- [5] Ma, C., Byun, B., Kim, I. and Lee, C.-H. "A detection-based approach to broadcast news video story segmentation" in Proc. of ICASSP 2009.
- [6] Barzilay, R., Collins, M., Hirschberg, J. and Wittaker, S. "The rules behind the roles: identifying speaker role in radio broadcast" in Proc. of 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence. AAAI Press/The MIT Press, 2000, pp. 679–684.
- [7] Liu, Y. "Initial study on automatic identification of speaker role in broadcast news speech," in Proc. of Human Language Technology Conference of the NAACL, 2006, pp. 81–84.
- [8] Vinciarelli, A. "Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling," in IEEE Transactions on Multimedia, vol. 9, no. 6, pp. 1215–1226, 2007.
- [9] Salamin, H., Favre, S. and Vinciarelli, A. "Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction," IEEE Transactions on Multimedia, vol. 11, no. 7, pp.1373–1380, 2009.
- [10] Bigot, B., Ferrané, I., Pinquier, J. and R. André-Obrecht, "Speaker role recognition to help spontaneous conversational speech detection," in Proc. of Int. Workshop on Searching Spontaneous Conversational Speech, pp. 5–10, 2010.
- [11] Yaman, S., Hakkani-Tür, D., and Tür, G. "Social role discovery from spoken language using dynamic bayesian networks," in Proc. of INTERSPEECH, 2010, pp. 2870–2873.
- [12] Bazillon, T., Maza, B., Rouvier, M., Bechet, F. and Nasr, A. "Speaker role recognition using question detection and characterization," in Proc. of INTERSPEECH, 2011, pp. 1333–1336.
- [13] Hutchinson, B., Zhang, B. and Ostendorf, M. "Unsupervised broadcast conversation speaker role labeling," in Proc. of IEEE ICASSP, 2010, pp. 5322 – 5325.
- [14] Damnati, G. and Charlet, D. "Robust speaker turn role labeling of TV broadcast news shows," in Proc. of IEEE ICASSP, 2011, pp. 5684 – 5687.
- [15] Dufour, R., Estève, Y. and Deléglise, P. "Investigation of spontaneous speech characterization applied to speaker role recognition," in Proc. of INTERSPEECH, 2011, pp. 917–920.
- [16] Salamin, H., Truong, K., Mohammadi, G. and Vinciarelli, A. "Automatic Role Recognition Based on Conversational and Prosodic Behaviour," in Proc. of ACM Multimedia, 2010, pp. 847–850.
- [17] D. Charlet, C. Barras and J.S. Lienard, "Impact of Overlapping Speech Detection on Speaker Diarization for Broadcast News and Debates", in Proc. of ICASSP, 2013.
- [18] Bigot, B., Senay, G., Linar'es, G., Fredouille, C. and Dufour, R. "Person Name Recognition in ASR outputs using Continuous Context Models" in Proc. of ICASSP, 2013.