Looking for relevant features for speaker role recognition

Benjamin Bigot, Julien Pinquier, Isabelle Ferrané, Régine André-Obrecht

IRIT - Université de Toulouse, France
118 route de Narbonne F-31062 Toulouse Cedex 9
{bigot, pinquier, ferrane, obrecht}@irit.fr

Abstract

When listening to foreign radio or TV programs we are able to pick up some information from the way people are interacting with each others and easily identify the most dominant speaker or the person who is interviewed. Our work relies on the existence of clues about speaker roles in acoustic and prosodic features, speaker segmentations, and content structure. When listening to foreign radio or TV programs we are able to pick up some information from the way people are interacting with each others and easily identify the most dominant speaker or the person who is interviewed. Such observations have been used in [1] to summarize broadcast news documents.

Speaker role recognition consists in assigning automatically a role to the speakers appearing in audiovisual data. This research topic holds an important place in the fields of Information Retrieval in spoken documents and social interaction analysis.

On 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. When listening to foreign radio or TV programs we are able to pick up some information from the way people are interacting with each others and easily identify the most dominant speaker or the person who is interviewed. Such observations have been used in [1] to summarize broadcast news documents.

Speaker role information is also reported as an important feature for topic indexing [2] and for story segmentation, relying either on the detection of the anchorman [3] or of the journalists [4]. In [5], speaker roles (e.g., project manager, marketing expert and others) are used to achieve a dialogue segmentation of meeting recordings taken from the AMI Meeting Corpus. Another work [6] concerns the detection of patient case discussions during medical team meetings based on the speaker medical role: radiologist, surgeon, oncologist and others. Different role definitions, input data, features, recognition processes are used or carried out to achieve this common goal.

Our ultimate purpose is to analyze speaker interactions and characterize the nature of the speech (prepared, conversational or spontaneous speech); such information will be introduced as prior knowledge in automatic speech recognition systems to choose the most appropriate language model (EPAC Project†). As a preprocessing to the interaction analysis, we propose to identify the role of each speaker, but this new context, and especially the necessary corpora independency, motivates us to revisit the state-of-the-art role recognition methods and to propose an alternative one.

Before describing our role recognition system in section 4, we give a brief overview of some role identification processes (section 2) and we describe the experimental corpus in section 3. Experiments and results are discussed in section 5.

2. State-of-the-art

Most of the work in speaker role recognition is based on a speaker diarization process of the audio flow and methods are evaluated on broadcast news data or homogeneous shows. In Barzilay’s work [7], automatic speech transcriptions and manual speaker boundaries are used as inputs and a speaker label (among anchor, journalist and guest speaker) is assigned to each speaker utterance. The role identification is guided by what the current or previous speaker said. This process, based on lexical features as well as speaker introduction phrases, is highly text dependent. The recognition is evaluated on a set of recordings of a same program.

In [8], two statistical approaches, Hidden Markov Models and maximum entropy, are proposed by Liu in order to attribute speaker roles chosen among three categories: anchor, reporter and other. For model training and for role identification, sets of human transcriptions and manual speaker boundaries are used. Like in the previous work, a decision is taken at the speaker utterance level.

More recently, Vinciarelli [9] has proposed methods for individual role recognition, applied after an automatic speaker diarization step. In this approach, the role assigned to each speaker characterizes rather his behaviour, that is his whole intervention and the way he interacts with others. This study is based on distributions of intervention durations and social network analysis.

These propositions have been followed up by Salamin’s contribution [10] in which an affiliation network is created, defining relationship between speakers and pertinent acoustic features extracted to characterize speaker interactions. A statistical approach is then applied to assign a role to a speaker. In contrast to most of the previous studies, mainly based on broadcast news and on quite homogeneous corpora (same program recordings), role detection is also applied to less structured content (meeting data).

As said in section 1, to be more independent to any transcription and generic enough to be applied to various audio contents (different type, structure, duration and even language), we developed a 3-role detection system on a heterogeneous corpus [11]. We used different sets of features assuming that clues about speaker roles can be found in prosodic, basic signal and

†http://epac.univ-lemans.fr/
Table 1: Performances of some state-of-the-art speaker role recognition systems on Broadcast News (BN) and/or Talk Shows (TS) - Performances are expressed in terms of % of well labelled: (1) segments, (2) speaker turns, (3) data time, (4) speakers.

Ref. | Corpus (Number Shows, Source Type/Volume) | Roles | Perf. |
--- | --- | --- | --- |
[7] | NIST TREC SDR (35), BN-17h | 3 | 80% (1) |
[8] | TDT4 (336), BN-170h | 3 | 80% (2) |
[9] | Swiss Radio (96), BN-25h | 6 | 85% (3) |
[10] | Swiss Radio (96), BN-23h | 6 | 82% (3) |
| Radio (27), TS-27h | 6 | 88% (3) |
| AMI Meeting Corp. (138)-45h | 4 | 46% (3) |
| 79% (3) |

temporal information extracted from audio signal and speaker segmentations. Each speaker was represented by a feature vector and supervised classification was done after a dimensionality reduction step. Performances, similar to the state-of-the-art (see Table 1), were obtained on a heterogeneous corpus.

In this paper, we investigated two extensions of this previous work: first we address now a 5-role problem by refining the previous 3 classes and secondly, we apply a feature selection method (instead of dimensionality reduction) before the classification step.

3. Corpus

For this work, we have used several documents taken from the development and test corpora of the ESTER2 evaluation campaign [12]. This corpus consists of 46 audio documents: 20 for development (DEV) and 26 for test (TEST) all recorded on 4 radio stations: TVME, Africa One, France Inter and RFI. Each document corresponds to one radio show: 41 news programs and 5 talk-shows. We observe a large diversity among them in terms of show duration (from 10 minutes up to 1 hour), time slots (morning or afternoon) and number of speakers (from 4 up to 21). All DEV documents (435 minutes) have been recorded in June and July 2007 and the TEST ones (440 minutes) in January and February 2008. Ground truth for speaker diarization was provided by the organizers of the ESTER2 campaign. Therefore, the quality of the automatic speaker diarization, used later as input, can be measured. A full description of these documents is available in [11].

We have created the speaker role reference on the manual speaker segmentations DEV-ref and TEST-ref with 3 "classical" roles: Anchor, Journalist and Other. Inside the classes Journalist and Other can be found Punctual and Non-punctual speakers. We define a Punctual speaker as a speaker who appears in only one segment. This distinction allows us to refine our 3 roles and to obtain 5 roles based on the following definitions:

- **Anchor** (noted A): the dominant speaker, appearing through the entire program, introducing news stories, guiding the program, animating the show and introducing other speakers.
- **Punctual Journalist** (noted PJ): a professional speaker working for the show, reporting a specific topic without interacting with any other speaker.
- **Journalist** (noted J): a non-punctual and professional speaker who may be involved in a conversation, an interview or be in charge of a chronicle.
- **Punctual Other** (noted PO): commonly heard during news stories, a man from the street or someone in a sound bite who does not interact with others.
- **Other** (noted O): a guest interviewed by a journalist or the anchorman, a person present in the studio or contacted by phone.

These definitions are generic enough to fit different broadcast programs. For the annotation of the speaker roles in TEST-auto, we only keep the automatic speaker segmentations which match significantly to the manual ones. The roles from TEST-ref are then reported to TEST-auto. Besides, we only take into account "significant" speakers (speaking more than 10 seconds). Evaluations are then conducted on the following populations:

- DEV-ref: 20 Anchor, 111 Journalist (including 39 PJ), 120 Other (including 55 PO),
- TEST-auto: 26 Anchor, 90 Journalist (including 35 PJ), 87 Other (including 28 PO).

4. Speaker role recognition system

4.1. System overview

The front-end of our system is a speaker diarization module. The tool we use to obtain this automatic segmentation was developed in our team by El Khoury [13]. The first step is based on the Generalized Likelihood Ratio (GLR) and the Bayesian Information Criterion (BIC) methods. With this tool, the overall speaker Diarization Error Rate (DER) on TEST-auto is equal to 7.82%.

Given these segments, the distinction between punctual and non-punctual speakers can be done. Punctual and non-punctual speaker segments are processed in two different ways both based on extracting and selecting features and then applying a classification step. The difference lies in the number of extracted features, the set of selected features and the classification method applied. Besides, the strategy we chose consists in solving a 2-class problem at each classification step. All this process is described in Figure 1.

After feature extraction, each speaker is then represented by one feature vector. In our previous work, an important set of features was used. Here, we propose to study the true impact of these features on role recognition. At each classification step, we proceed to feature selection to keep the most relevant features. After describing the initial set of features, we describe how feature selection and classification are carried out.

4.2. Feature extraction

Two types of low-level features are extracted from the segments of each speaker: 24 temporal and acoustic features and 12 prosodic ones. Each speaker is represented by a 36-feature vector.

4.2.1. Temporal and Acoustic features

From previous work [14], we concluded that clues about speaker roles are carried by the temporal arrangement of speaker diarization results. This leads to a first temporal feature subset.
It is composed of the number of segments, the length of segments and intersegments (between two consecutive segments of the same speaker) through their mean, variance, maximum, minimum, the overall speech activity and the span of the speaker intervention (from the beginning of his first segment to the end of his last one). These features are combined to compute the speaker inactivity rate, the ratio between the number of segments and the overall activity and the ratio between the number of segments and the speaker intervention span.

A second feature subset, the acoustic subset gathers the audio signal power corresponding to each speaker segmentation. The signal power is compared to a given threshold equal to 15% of the average power value of the segment set. By this way high energy zones can be distinguished from low energy ones. Finally, 10 features are calculated over the whole signal (mean and variance of the signal power) and over the high (resp. low) energy zones only (mean, variance, maximum and minimum of the signal power values).

4.2.2. Prosodic features

The prosodic feature set contains 12 terms. Four of them are computed from the fundamental frequency estimation [15]: mean, variance and maximum pitch value and the average number of voiced zones per second. The other eight features are calculated using the vocalic segmentation of [16]. This segmentation provides an estimation of the positions and durations of vocalic segments and silences. From the whole set of vocalic segments (and respectively of silence segments) corresponding to the speaker, we compute their overall number, their average number per second, the mean and variance of their duration.

4.3. Feature selection and classification methods

We investigate supervised methods: Gaussian Mixture Models (GMM), K-nearest neighbours algorithm (K-nn) and Support Vector Machines (SVM) in its linear and gaussian kernel versions. Before every classification step, we perform a Sequential Backward Feature Selection (SBFS) [17]. The SBFS is conducted on training samples using a leave-one-out cross validation. This algorithm starts with the full feature set and one feature is removed at each iteration. The most discriminative features are kept regarding to the accuracy. Classification is then conducted on the test set using the same recognition method with this best feature set. The characteristics of each method (K for K-nn, number of GMM and SVM parameters) are estimated using a leave-one-out on the training samples.

5. Experiments and results

In our experiments, the DEV-ref corpus is used for training and TEST-auto for evaluation. Performances are expressed in terms of accuracy with a 95% confidence interval. They give the proportion of speakers whose role has been correctly recognized over the overall number of tested speakers. The SBFS method keeps the most discriminative features which are highly specific to the DEV-ref corpus. Role recognition is performed with these features applied to the corpus TEST-auto. Table 2 contains, for each role recognition configuration, the number of features kept over the overall number of tested speakers. The SBFS method keeps the most discriminative features which are highly specific to the DEV-ref corpus. Role recognition is performed with these features applied to the corpus TEST-auto. Table 2 contains, for each role recognition configuration, the number of features kept over the overall number of tested speakers. The SBFS method keeps the most discriminative features which are highly specific to the DEV-ref corpus. Role recognition is performed with these features applied to the corpus TEST-auto. Table 2 contains, for each role recognition configuration, the number of features kept over the overall number of tested speakers. The SBFS method keeps the most discriminative features which are highly specific to the DEV-ref corpus. Role recognition is performed with these features applied to the corpus TEST-auto.

![Figure 1: System Overview.](image)

Table 2: Feature selection performances on TEST auto.

<table>
<thead>
<tr>
<th>Anchor/Non-Anchor</th>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM (1 gaussian)</td>
<td>36</td>
<td>96.4 ± 3.1</td>
</tr>
<tr>
<td>K-nn (k=3)</td>
<td>8</td>
<td>86.4 ± 5.7</td>
</tr>
<tr>
<td>SVM rbf</td>
<td>nc</td>
<td>nc</td>
</tr>
<tr>
<td>SVM linear</td>
<td>14</td>
<td>92.1 ± 4.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>J/O</th>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM (1 gaussian)</td>
<td>34</td>
<td>68.4 ± 8.6</td>
</tr>
<tr>
<td>K-nn (k=17)</td>
<td>22</td>
<td>68.4 ± 8.6</td>
</tr>
<tr>
<td>SVM rbf</td>
<td>30</td>
<td>63.2 ± 8.9</td>
</tr>
<tr>
<td>SVM linear</td>
<td>26</td>
<td>74.6 ± 8.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PJ/PO</th>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM (1 gaussian)</td>
<td>20</td>
<td>80.9 ± 9.8</td>
</tr>
<tr>
<td>K-nn (k=13)</td>
<td>9</td>
<td>49.2 ± 12.4</td>
</tr>
<tr>
<td>SVM rbf</td>
<td>21</td>
<td>85.7 ± 14.7</td>
</tr>
<tr>
<td>SVM linear</td>
<td>10</td>
<td>88.9 ± 7.8</td>
</tr>
</tbody>
</table>

The feature selection is performed using the K-nn classifier is reduced from 36 down to 8 and the recognition reaches an accuracy of 86% on TEST-auto. The best performance 96% is obtained with a single Gaussian Model for which the 36 features were selected. Finally, the linear SVM is an acceptable compromise with an accuracy of 92% (3 mis-classified speakers comparing to the Gaussian model system); the feature number is reduced to 14, two features are related to the intersegments (variance of the length and minimum length), 3 to the segment length (minimum, maximum and variance), 3 to the power signal over high activity zones (variance, maximum and mean value), and 6 prosodic features (maximum value of the pitch, number of vowels, variance on the vowel length, number of silences, rate of silences and average duration of silences). Anchor is well recognized among the non-punctual speakers thanks to temporal and prosodic features in an equivalent proportion and only with a third of the original feature set. Four of them (the variance of the power signal on high activity zones, the maximum pitch value, the number of silences and the average duration of silences) are found with each of the four classification methods.
Classifying Journalist/Other is a more difficult task. At first, the number of features kept is quite high (higher than 22). Secondly, the best accuracy reached is only 74% with a linear SVM. Among the 26 features kept for this classifier, 18 belong to the temporal feature set while 8 to the prosodic one (pitch average and maximum, voiced zone rate, number and average duration of vowels and silences, mean and variance of the silence duration). Most errors consist of Other miss-classified as Journalist: these Other are found as non-punctual speakers in the TEST-auto as they are actually punctual speakers in the reference. This is a consequence of the speaker diarization system errors.

For the third classification step Punctual Journalist/Punctual Other, the feature selection coupled to the K-nn classifier appears to keep too specific features with only 9 features; the accuracy is about 49%. The best performances, in terms of accuracy and number of features are obtained using the linear SVM classifier: accuracy of 88.9% with 10 features selected, 4 temporal and acoustic features (overall speaker activity, the variance of power signal over the whole speaker segmentation, the maximum and minimum values of the power signal on the high activity zones) and 6 prosodic features (maximum value of the pitch, average voiced zone rate, overall number of vowels and silences, the mean and variance of the silence duration). Without neglecting the influence of temporal features, we observe that almost all the prosodic features have been kept by the SBFS. An interpretation may be that usually Punctual Journalist speakers prepare their speech and Punctual Other ones speak more spontaneously.

We have conducted the role recognition with the best feature sets and classifiers; the configurations are indicated in bold in Table 2. The overall role recognition accuracy reaches 85.6% ± 8% in terms of speaker roles well recognized. This result has to be compared with former investigations conducted in [11] where a classical PCA was applied on the feature set and where the best system reached an accuracy of 72%. The gain of 13% is obtained thanks to a hierarchical classification approach and an efficient feature selection step. In future works we will start exploiting these results for analyzing speaker interactions in a content structuring task.

6. Conclusion

This paper describes our contribution to the domain of speaker role recognition. Our work relies on the assumption of the existence of clues about speaker roles in acoustic, temporal and prosodic features extracted from audio files and from speaker segmentations.

Each speaker is represented by a 36-feature vector. In contrast to most existing off-the-shelf approaches we do not use the structure of the document to recognize the roles of the interlocutors and we do not use the results of an automatic speech recognition system as an input.

We classify five roles (instead of three roles in our previous work) occurring in broadcast news: Anchor, Journalist, Other, Punctual Journalist and Punctual Other. A punctual speaker appears in only one segment. Punctual and non-punctual are now treated separately and hierarchically. We also investigate the relevance of each feature by using a sequential backward feature selection method.

Results highlight that low-level features, robust to speaker diarization errors, enable role recognition in heterogeneous content. Anchor is well recognized with a small number of temporal and prosodic features. Classifying non-punctual Journalist and non-punctual Other requires much more features. The Punctual Journalist and Punctual Other classification relies on a small number of prosodic features. Experiments have been conducted over automatic speaker segmentations and the accuracy reaches 85% of the speaker roles correctly labelled.

7. Acknowledgements

This work is in the Project ANR-06-CIS6-MDCA-006.

8. References