IDIAP RESEARCH REPORT

Role Recognition in Broadcast News Using Social Network Analysis and Duration Distribution Modeling

A.Vinciarelli [a] [b]

IDIAP–RR 06-35

March 2007

to appear in
IEEE Transactions on Multimedia

[a] IDIAP Research Institute- vincia@idiap.ch
[b] Ecole Polytechnique Fédérale de Lausanne (EPFL) - 1015 Lausanne (Switzerland)

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 06-35

Role Recognition in Broadcast News Using Social Network Analysis and Duration Distribution Modeling

A.Vinciarelli

March 2007

to appear in
IEEE Transactions on Multimedia

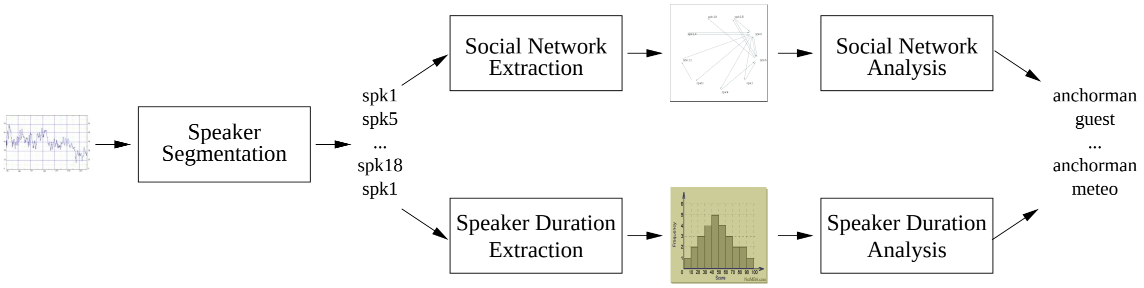

Figure 1: Overall approach. The speaker segmentation obtained after the first stage of the process is used to extract Social Networks and duration distributions. These are then analyzed in order to perform role recognition.

1 Introduction

Many approaches in the multimedia domain aim at extracting *high level information* from different sources (videos, audio, etc.), i.e. information which is not explicitly stated in the data, but it can be extracted through an automatic process. This kind of works include (the list is not exhaustive) event detection [24], speaker diarization [2], action recognition [20] etc. (see Section 2 for a short survey). In most cases, the information is used for indexing purposes, i.e. to support applications such as Information Retrieval, browsing, etc. which enable a more effective access to the data content [16].

This work focuses on an information that, to our knowledge, has never been addressed before: the individuals *role*. The reason is that, in several cases, people in multiparty recordings (i.e. recordings involving several individuals) play a specific role, i.e. they follow a more or less rigorous plan that imposes constraints on frequency and timing of their interventions. This is the case of radio and television programs that are based on a *format* including specific roles for each participant. This work shows how such a role can be recognized by applying Social Network Analysis (SNA) [26], the domain studying the interaction between people in social environments, or by analyzing the duration distribution of speakers interventions. The experiments performed in this work, focus on a specific kind of data, namely radio news bulletins where each speaker plays one among six predefinite roles.

The approach we propose is illustrated in Figure 1. Audio recordings are first segmented into single speaker segments using a system based on HMMs [15] and Poisson Stochastic Processes [21]. The speakers are not known a-priori and they are labeled with a code that is not related to their identity. The resulting segmentation is thus a sequence of speaker IDs that can be used to extract a Social Network or the time distribution across different speakers. Such information is then used to perform role recognition.

The main characteristic of the SNA approach is that it takes into account only *relational data* (see Section 5 for more details) and it is independent of speakers identity and recording length. The only important aspect is the interaction pattern between different speakers and not who they are and how long they talk. This can be an advantage under two main respects: the first is that, in many cases, the format is the same, but the speakers (or at least part of them) change at each recording. This is the case in our data where roughly 50 percent of the material contains speakers that appear only once and, on average, there are 4 new speakers per recording. The second is that there are speakers that play different roles in different recordings. This happens, in the case of our data, for around ten frequent speakers that account for roughly 50 percent of the corpus material.

The approach based on the duration distributions considers the fraction of the recording time each speaker accounts for. The value of such a fraction distributes differently depending on the role and this provides the information necessary to assign the speakers their role (see Section 6 for more details). The main disadvantage of this approach is that it can distinguish between different roles only when the distribution of the respective durations is not overlapping too much or the a-priori probabilities of the roles are different. On the other hand, the approach is independent of the identity of the speakers and can thus deal with data where the same role is played by different speakers or the same speaker plays different roles.

The above approaches rely on different information sources and are thus expected to be *diverse*, i.e. to make different errors over the same data. For this reason, the two approaches are combined to verify whether the role recognition performance can be improved.

An important advantage of the process depicted in Figure 1 is that it captures the *authors perspective* [25]. In fact, different roles can be thought of as different parts of a text (e.g. chapters, sections, etc.) because they correspond to different functions in delivering information. For this reason, role recognition enables one to capture the structure given to the recordings by data producers. On the other hand, while roles are well defined in data created in a *production environment* (e.g. news, movies, documentaries, etc.), they are less evident, or even absent, in data collected in more spontaneous environments like meeting recordings and home made movies. This can limit the application domain of role recognition techniques.

To our knowledge, the role recognition problem has been addressed only in few works (see Section 2) though it could improve several applications. Speakers role can enhance browsers (users can access specific data segments based on role), summarization systems (segments corresponding to certain roles can be retained in the summary as more representative of the content than others), thematic segmentation approaches (specific roles are often related to specific topics), etc.

The rest of this paper is organized as follows: Section 2 presents a survey of works dedicated to the extraction of information from audio recordings, Section 3 describes data and roles, Section 4 shows our speaker segmentation approach, Section 5 introduces Social Network Analysis, Section 6 presents the duration distribution based approach, Section 7 presents the combination technique, Section 8 presents experiments and results and Section 9 reports conclusions.

2 Previous Work

To our knowledge, the role recognition problem has been addressed only in few works: an approach based on lexical specificities is proposed in [5] to recognize the roles of speakers in broadcast news, and a technique for the classification of TV shows segments into *host* and *guest* is described in [14]. The first work uses the audio, while the second takes into account the visual channel of videos. Role recognition is one of the many approaches trying to extract *high level* information from multimedia recordings, i.e. to extract information which is not explicitly stated in the data and it rather requires an *abstraction* process. For this reason, this section presents a survey of the works extracting different kinds of information from audio recordings.

The works dedicated to audio archives can be roughly divided into two groups. The first includes the works where spoken documents are first converted into texts through Automatic Speech Recognition (ASR) [15] and then indexed with approaches developed for digital texts [4] (see [18] for an extensive survey on this approach). The second includes works that consider audio far richer than a simple textual transcription [16], [25] and try to extract information like speaker identity and emotional state as well as affective content and different kinds of events. This survey focuses on the latter aspect.

Several works have been dedicated to the detection of events considered of particular importance in applications like browsing, summarization and indexing. Such an approach is especially common in analyzing sport videos where certain events (e.g. goals or red cards in soccer) are likely to influence the final result and are thus considered as the highlights of the whole match [24], [3], [8]. Most of the approaches are based on models of specific targeted events. In [24], [3] the authors identify important moments by detecting cheering and excited speech as well as information extracted from the visual channel. Experiments performed on several hours of recordings show that events like goals or yellow cards in soccer matchs can be detected with satisfactory performance. In [8], the detection

Role	AM	SA	GT	AB	MT	IP
Fraction (%)	41.2	5.5	34.8	7.1	6.3	4.0

Table 1: Corpus Characteristics. This table reports the percentage of corpus time that each role accounts for.

of important events is used to create summaries containing only the match highlights and accounting for less than 5 percent of the original duration. Audio based information is very important also in the analysis of meeting recordings where speech is the main channel of communication between the various participants. Several works [20], [28], [10] are aimed at identifying group actions like discussions, agreement and monologues. The proposed approach is based on machine learning algorithms fed with features (e.g. speaker location) extracted from both audio and visual channel.

Other works [27], [17], [29] focus on information that can be extracted from audio without trying to model higher level events. In [27], the authors address the problem of detecting crosstalk, i.e. the presence of other voices in the lapel microphone of a certain speaker. This is necessary in order to enhance the performance of speech recognizers, but also to detect moments when several persons talk together (an information important for the dynamics of the meeting). The subject of [17] is the detection of speaker emotional states. The authors perform experiments on a corpus of conversations collected at a call center. The goal is to detect unsatisfied or stressed users in order to activate help mechanisms. The work presented in [29] shows how audio can be segmented into classes like music, speech and singing. that can be helpful to segment videos.

Some works address the above problem by modeling the reaction (in both affective and attentional terms) that certain characteristics of the audio and visual channels (e.g. high energy sounds and fast moving images) are likely to induce in the audience. The advantage of such an approach is that it can lead to general models valid over a wide range of data. In [19], the authors use a model of human attention in order to detect the events that are more likely to attract the audience and that must thus be selected when creating a video summary. Authors observe that humans tend to pay more attention to speech and music, thus detect this kind of information. Experiments performed over five videos (around 70 minutes in total) show that satisfactory results (as evaluated by 20 human assessors) can be achieved. A similar perspective is used in [12], [11] where important events are associated with audio characteristics likely to produce, following psychophisiologists, excitement in the listeners. Experiments performed over a database of soccer matchs and Hollywood movies (for a total of few hours) show that segments corresponding to the excitement model contain important events in both kinds of data. A similar approach is used in [7] to perform affect-based indexing and retrieval of films. Since movie directors use loud music or noise to generate emotional reactions in the audience, highly emotional scenes are found by detecting energy peaks in the audio signal.

3 Data and Roles

The experiments of this work have been performed over a corpus of 96 news bulletins collected during february 2005 at Radio Suisse Romande (RSR), the Swiss national broadcasting service. They represent the whole set of bulletins broadcasted during february 2005 working days (during the week ends RSR diffuses only short communicates). The corpus is thus realistic and representative of this specific kind of radio news. The total duration of the data is 18 hours, 56 minutes and 23 seconds and the average bulletin length is 11 minutes and 50 seconds. The shortest recording is 9 minutes and 4 seconds long, while the longest one lasts 14 minutes and 28 seconds. The standard deviation (1 minute and 17 seconds) accounts for the high variability of different recording lengths. The average number of speakers is 11.0 with a standard deviation of 1.6. The speakers account for 99 percent of the data and the remaining 1 percent corresponds to music, noise, jingles, etc.

In our experiments, we identified six roles that can label any speaker. In other words, each speaker can be assigned one of the six roles and no speaker must be left unlabeled because no role is available. The first is the *anchorman* (AM), i.e. the role of the persons coordinating the bulletin, introducing other speakers and, in some cases, discussing the most important issues. The main characteristics of AMs are that they interact with most of the other speakers and they appear all allong the recordings. In our data, the identity of the AM changes at each bulletin and there are around 10 persons playing alternatively the AM. The second role is the *secondary anchorman* (SA), i.e. the person that supports the AM by announcing new topics. The main characteristic of the SAs is that they interact mostly with the AM and their interventions are typically short (20-25 seconds). The SA role is played by the same group of persons playing the AM. At each bulletin, the identity of the SA changes and the same person can play, in different bulletins, both SA and AM. The third role is the *guest* (GT), i.e. a person invited to express an opinion or to report about a single and specific topic. GTs appear only once in a given bulletin (but they can appear in several bulletins) and they interact with no more than two persons. The GTs change everyday and only few of them (in general the same journalists that in other bulletins play the AM role) appear in different days. The fourth role is the *interview participant* (IP), i.e. someone who is involved in an interview where two or more persons have an exchange and interact with each other. Like in the case of GTs, the identity of IPs changes everyday and only few of them appear more than once in the data corpus. The last two roles are *abstract* (AB) and *meteo* (MT). ABs appear at the beginning of the bulletins and provide a summary. Sometimes the AB role is played by the AMs, but in other cases there is a specific speaker. The identity variability is the same as in the case of AMs, i.e. relatively few persons appear alternatively as AB in different bulletins. The MTs appear typically at the end of the news and they give the wheather forecast. This is the only role which is played by few persons (two in the case of our data). Since they appear at an extreme of the newscast, both MT and AB interact with no more than one person. Table 1 shows the percentage of the total time represented by each role. AM, SA, AB and MT are played by a single person and are represented in every bulletin, while GT and IP are played by a number of persons which is not known in advance and are not necessarily represented in every bulletin.

4 Speaker Clustering

Given a multiparty recording of duration T, the goal of a speaker segmentation algorithm is to find a sequence $S = \{(s_1, \tau_1), \ldots, (s_N, \tau_N)\}$ of pairs (s_i, τ_i) representing segments where speaker and duration are s_i and τ_i respectively. The number of speakers G and turns N is not known a priori and $\sum_{i=1}^{N} \tau_i = T$. Such a problem can be solved through a speaker clustering algorithm, i.e. an unsupervised technique capable of grouping the feature vectors extracted from the recordings so that each cluster corresponds to a single speaker. The clustering approach applied in this work (see [2], [1] for a detailed description) is based on ergodic continuous density Hidden Markov Models (HMM) with Gaussian Mixture Models (GMM) as emission probabilities [15]. Each cluster corresponds not only to the observation vectors (see below) belonging to a certain speaker, but also to a state in the HMM and to its emission probability function.

The first step of the SC process is the conversion of the audio data into a feature vectors sequence $O = \{o_1, \ldots, o_K\}$. Several feature extraction techniques are available in the literature (see [13] for an extensive survey), but this work uses 12 dimensional *Mel Frequency Cepstral Coefficients* (MFCC) vectors extracted every 10 ms from a 30 ms long window [13]. The reason is that MFCC features have, on average, higher performance in speaker recognition tasks (they are thus effective in capturing speaker voice characteristics) and extensive experiments show that they lead to better results in speaker clustering experiments [1].

Once the sequence O is available, it is possible to initialize the ergodic HMM (see above). Since G is not known a-priori, it is necessary to make an initial guess $G^{(0)}$ that must be significantly higher than the actual value of G. The initial clusters are obtained by simply segmenting O into $G^{(0)}$ uniform segments. The vectors in each segment are assumed to belong to the same cluster and they are used to train, using the Maximum Likelihood algorithm (ML) [15], a GMM that will be the emission

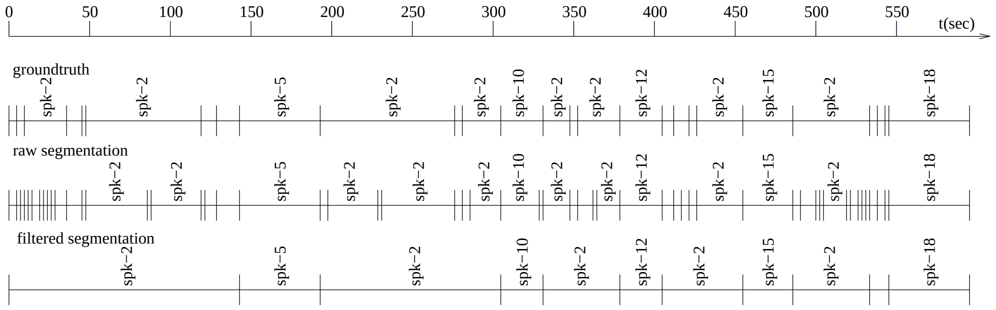

Figure 2: Speaker Segmentation. This figure shows the distribution of speaker changes in groundtruth (upper plot), raw (central plot) and filtered (lower plot) speaker segmentation. The label is missing for shorter segments only because there is not enough space. The system provides a label for each segment and 99% of the material corresponds to speakers talking.

probability function of the HMM. The HMM could be trained as a whole using the Baum-Welch algorithm [15], but separate training of each GMM is computationally less expensive and leads to similar results. The only problem is that it is not possible to train the transition probabilities, but these play no role in a speaker segmentation problem and can be set to a uniform distribution [1].

The result of the training is an HMM with parameter set $\Theta^{(0)}$ that can be aligned with O using the Viterbi algorithm [15] to obtain the best sequence of states (i.e. speakers):

$$q^{(0)} = \arg\max_q p(q|O, \Theta^{(0)}) \tag{1}$$

where q is a sequence of states. As a result of the alignment, the data assigned to each cluster are changed with respect to the initialization and it is possible to re-train the GMMs of each cluster to obtain a new HMM with parameter set $\Theta^{(1)}$:

$$\Theta^{(1)} = \arg\max_\Theta p(q^{(0)}|O, \Theta) \tag{2}$$

where $\Theta = \{\theta_1, \ldots, \theta_{G^{(0)}}\}$, i.e. the parameter set of the HMM can be thought of as a set of GMM parameters.

Since the number of speakers $G^{(0)}$ is higher than G, the data is oversegmented and there are clusters that should be merged since they contain data belonging to the same speaker. For this reason, two states are merged when the following condition is met:

$$\log p(O_{m+n}|\theta_{m+n}) \geq \log p(O_m|\theta_m) + \log p(O_n|\theta_n) \tag{3}$$

where O_m, O_n and O_{m+n} are the observation vectors attributed to cluster m, n and their union respectively, θ_m and θ_n are the parameters of GMMs in states m and n and θ_{m+n} are the parameters of a GMM trained with EM on O_{m+n}.

After the merging process, the resulting HMM (that has less states than the original one) is aligned with O and the parameters of the GMMs are trained as explained above. At the i^{th} iteration, this leads to the state sequence $q^{(i)}$ such that:

$$q^{(i)} = \arg\max_q p(q|O, \Theta^{(i)}). \tag{4}$$

The main problem of the above criterion is that the likelihood of an HMM always decreases after reducing the number of parameters (this is always the case after merging two states) and this makes it impossible to find a stopping criterion that does not require a manually set threshold. A solution is to set $|\theta_{m+n}| = |\theta_m| + |\theta_n|$ when merging states m and n ($|\Theta|$ is the number of parameters in Θ). In this way, the number of parameters remains constant across different iterations and there is empirical evidence that the likelihood increases up to a certain point (until the merging involves states actually corresponding to a given speaker) and then it starts to decrease (when the merging involves states corresponding to different speakers). Although such an effect is not proved theoretically, there is empirical evidence that it leads to satisfactory results not only in this work, but also in other works presented in the literature [2].

4.1 Speaker Segmentation Smoothing

The result of the speaker clustering process described in the previous section is a sequence $S = \{(s_1, \tau_1), \ldots, (s_N, \tau_N)\}$ such that $\sum_{i=1}^{N} \tau_i = T$, where T is the total duration of the audio recording under examination (see Figure 2). Because of the high variability in the data, S contains many spurious turns that are not determined by an actual change of speaker, but rather by effects like background noise, music, crosstalk, etc. Figure 2 (central segmentation) shows that such an effect is particularly evident in certain points (e.g. at the beginning and between 500 and 550 seconds). This is an important problem not only because it decreases the quality of the speaker segmentation, but also because it introduces a high number of spurious interactions that heavily affect the Social Networks extracted from the data (see Section 5 for more details). It is thus necessary to post-process S in order to reduce as much as possible the amount of false interactions.

In order to address the above problem, we consider the groundtruth speaker segmentations of the data used in our experiments. Given a groundtruth segmentation $S^* = \{(s_1^*, \tau_1^*), \ldots, (s_K^*, \tau_K^*)\}$, there are $K-1$ speaker changes and the i^{th} one of them takes place at time $t_i^* = \sum_{n=1}^{i-1} \tau_n^*$. For each recording m it is thus possible to obtain a function $f_m(t)$ giving, at each time t, the number of speaker changes that took place in the time interval $[0, t]$. Since such functions increase by one at each time t_i^*, they are called *staircase* functions. If a recording corpus contains M multiparty recordings, the average number $n(t)$ of speaker changes at time t can be estimated as follows:

$$n(t) = \frac{1}{M} \sum_{m=1}^{M} f_m(t). \tag{5}$$

Figure 3 shows that the values of $n(t)$ can be fitted by a line of slope λ, in other words:

$$n(t) \simeq \lambda t. \tag{6}$$

Equation 6 means that the speaker change distribution in time is governed, like many other natural and technological phenomena, by a Poisson Stochastic Process (PSP) [21] and this can be used to detect spurious turns. In fact, if the t_is are distributed following a PSP it can be demonstrated that:

$$p(t_{i+1} - t_i < t) = 1 - e^{-\lambda t}. \tag{7}$$

where $t_{i+1} - t_i$ is the duration τ_i of the i^{th} segment. The odd of the probability in Equation 7 can be used to identify as spurious all the turns such that:

$$\frac{p(\tau < t)}{1 - p(\tau < t)} = \frac{1 - e^{-\lambda t}}{e^{-\lambda t}} \geq \frac{1}{2}. \tag{8}$$

In other words, the above expression states that a segment is considered as spurious when the probability of a segment being longer than τ is more than two times higher than the probability of being shorter than τ. Such a threshold has been fixed arbitrarily a-priori and no other values have been tried.

The last equation enables one to label each segment as spurious or non-spurious and along the segment sequence there are two possibilities: the first is that a spurious segment is between two non-spurious segments, the second is that a spurious segment neighbors, at least on one side, another

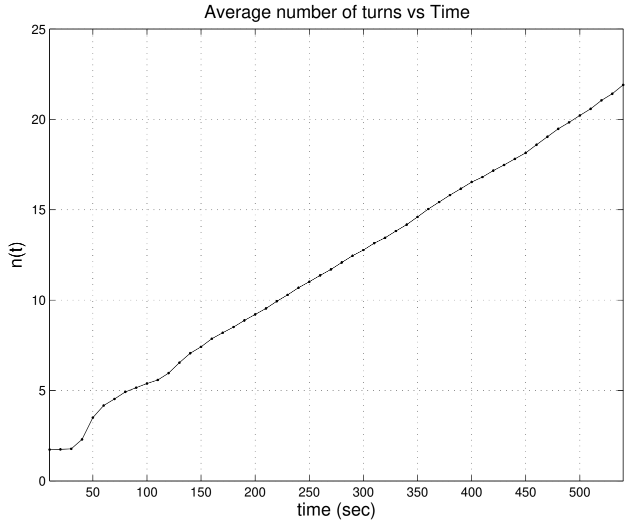

Figure 3: Poisson Stochastic Process. This plot shows the average number of turns at time t. The value of $n(t)$ is estimated every 10 seconds.

spurious segment. In the first case, the spurious segment is removed by attributing left and right halves to left and right neighbours respectively. In the second case all adjacent spurious segments are aggregated and the resulting segment is attributed to the most represented (in terms of time) speaker appearing in it. The resulting segment is always in the first situation described above, i.e. it is between two non-spurious segments. In fact, the aggregation starts at the leftmost spurious segment and ends at the rightmost spurious segment, then on the left and right side of the aggregation there are always non-spurious segments. If the segment resulting from the aggregation is still spurious following Equation 8, then it is attributed to left and right neighbours like in the case of spurious segments adjacent to non-spurious ones. Otherwise it is retained as a non-spurious segment.

Before applying the above smoothing algorithm, it is necessary to estimate the λ parameter. In order to perform experiments over the whole corpus at disposition, we used a leave-one-out approach, i.e. we estimated λ separately for each recording using the other data of the corpus. In this way, the separation between training and test set is guaranteed and the system is not fitted to the test data. The estimation is performed by fitting the points of Figure 3 with a first degree polynomial with the Least Square Method and by retaining the coefficient of the first power of the variable as λ estimate.

The application of the above algorithm removes most of the spurious segments, but it affects also some actual speaker changes (this is often the case of short questions in interviews). On the other hand, most of the spurious interactions affecting negatively the Social Networks extracted from the data are removed and this improves the results obtained through SNA (see Section 8).

5 Social Network Analysis

Social Network Analysis is the sociological domain studying the interaction between people in social environments [26], [23]. Given a group of individuals (often called *actors*), SNA focuses on the so-called *relational data*, i.e. all evidences of the fact that two or more persons interact with each other. At the same time, any *attribute data*, i.e. any characteristic specific of single individuals, is neglected. This is important because it enables our approach to be independent of speakers identity and time duration of segments and recordings.

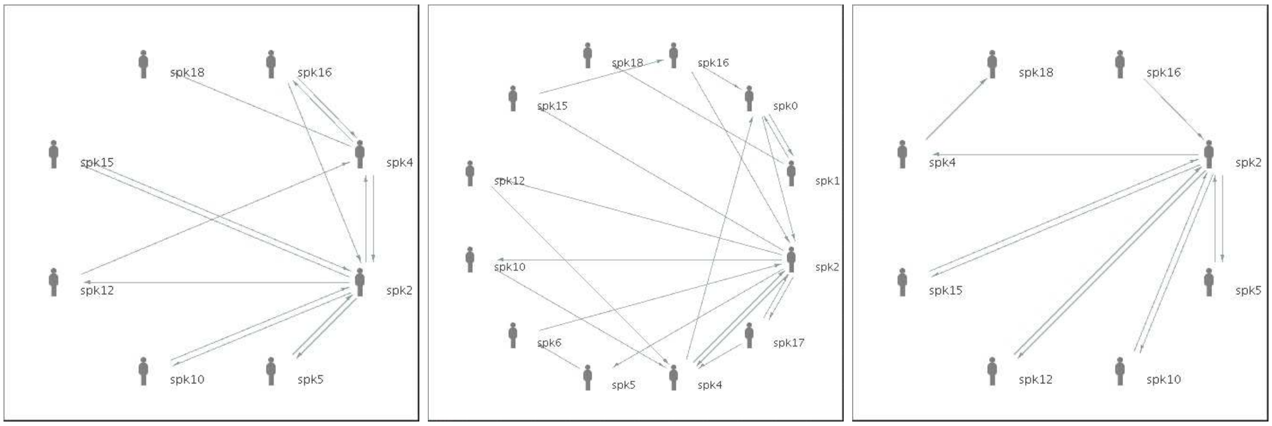

Figure 4: Social Networks. The networks shown in this figure correspond to the segmentations shown in the previous figure. The leftmost network has been extracted from the groundtruth, the central from the raw segmentation and the rightmost from the filtered segmentation.

The next subsections show how the relational data are extracted and how they are used to recognize the speakers role.

5.1 Social Network Extraction

The relational data are collected from the speaker segmentations obtained using the system described in Section 4. Following an experimental psychology approach [9], we use as evidence of the relationship between two individuals the fact that one of them talks at least once immediately before the other one in the recording. More formally, given the set $A = \{a_1, \dots, a_G\}$ of the G speakers involved in a recording and a relationship $R(a_i, a_j) : A \times A \rightarrow \{0, 1\}$, we state that $R(a_i, a_j) = 1$ if and only if speaker a_i talks immediately before speaker a_j in the speaker sequence obtained after the segmentation at least once. Since the order is taken into account, R is not symmetric, thus $R(a_i, a_j) = 1$ does not necessarily imply that $R(a_j, a_i) = 1$. The relational data is the set of ordered pairs (a_i, a_j) such that $R(a_i, a_j) = 1$.

Given a set of relational data, it is possible to create the so-called *sociomatrix* X where $x_{ij} = R(a_i, a_j)$ and to build the Social Network, i.e. the graph where each node represents an actor a_i and each edge represents a relationship. When the network is built using a non symmetric relationship, the corresponding graph is directed: an arrow from a_i to a_j means that $R(a_i, a_j) = 1$. Figure 4 shows the three networks extracted from the segmentations of Figure 2. While the networks have essentially a visualization purpose, the sociomatrices are used to perform the network analysis.

Figure 4 shows that not all of the speakers are connected directly, but they can be connected through *paths* passing through other speakers. By path it is meant a sequence of alternating nodes and arcs so that each arc has its origin at the previous node and its terminus at the subsequent node (nodes and arcs must be included in the sequence only once). The length d_{ij} of the shortest path going from a_i to a_j, i.e. the number of arcs included in the list, is often used as a measure of the distance between nodes and it plays an important role in actor centrality measures (see next section).

The distance between nodes can be calculated using the sociomatrix. The element ij of the X^2 matrix is calculated as $\sum_{k=1}^{G} x_{ik}x_{kj}$. Such a sum is different from zero only if there is at least one value of k for which both x_{ik} and x_{kj} are different from zero, in other words only if there is a path of length 2 connecting a_i and a_j and passing through a_k. Since the sociomatrix elements can be equal only to 0 or 1, the element $(X^2)_{ij}$ is the number of paths of length 2 connecting a_i and a_j. The same considerations apply to any power p of X and the distance between a_i and a_j can be calculated as follows:

$$d_{ij} = \min(\infty, \arg\min_p p : (X^p)_{ij} > 0), \tag{9}$$

i.e. the smallest value of p for which the element $(X^p)_{ij}$ is different from zero. The value of p cannot be higher than $G-1$, thus if $(X^p)_{ij} = 0$ for all values of $p \in \{1, \ldots, G-1\}$, the distance is set to infinity and the graph is said *disconnected*.

5.2 SNA Based Role recognition

The role recognition process consists in assigning each actor a_i a role $r_i = r(a_i)$. From a probabilistic point of view, the problem can be thought of as finding the roles r_i^* that maximize the following probability:

$$\vec{r}^* = \arg\max_{\vec{r}} p(r(a_i) = r_i, i = 1, \ldots, G) \tag{10}$$

where $\vec{r} = (r_1, \ldots, r_G)$. The above probability can be factored as follows:

$$\begin{aligned} &p(r_i = AM)p(r_j = SA|r_i = AM)\cdot \\ &\cdot p(r_k, k \neq i, j|r_i = AM, r_j = SA) \end{aligned} \tag{11}$$

and the last term of the product can be further factored:

$$p(r_k, k \neq i, j) = \prod_{k \neq i,j} p(r_k|r_i = AM, r_j = SA) \tag{12}$$

where we make the assumption that the roles played by actors different from SA or AM are statistically independent.

The rationale behind the above factorization is that AM and SA are represented in every bulletin and are played by a single person. For this reason such roles are assigned to the individual most likely to play them. On the contrary, GT and IP are not always represented and they can be played by an arbitrary number of persons, then they are assigned when they are the most likely role for a given speaker. Moreover, while the AMs can be recognized based on their *centrality* (see below), i.e. a sociometric characteristic associated to each individual, the other roles can be recognized only through the *fraction of interaction* they have with AM and SA (see below), i.e. a characteristic of a pair of individuals that can be used only when the AM is known. This requires to perform the recognition in a sequential way, i.e. by recognizing AM and SA first, and after the other roles (this is the reason of the factorization in Equation 11).

The factorization in Equation 11 corresponds to the simplifying assumtpion that the different terms are statistically independent. In this way, the value of the product can be found by maximizing separately the different terms and the resulting vector $\vec{r}^*$ is actually the one maximizing $p(\vec{r})$. One limit of such an approach is that the eventuality that the AM or the SA are split into more than one cluster is not taken into account. In fact, each cluster is assumed to correspond to one and only one speaker and the approach works only to the extent that such an assumption is true. In the case of our data, this does not create problems (see Section 8) because the AM voices are sufficiently represented to be never split into more than one clusters (only less represented voices tend to be split into several clusters).

The problem is now to model the different terms of Equation 11 given that the only available information is a set of labeled recordings and the set $\mathcal{X} = \{X_1, \ldots, X_N\}$ of their corresponding sociomatrices. Given a sociomatrix, there are two measures characterizing an actor a_i: the first is the *centrality* and the second is the *relative interaction* with another actor a_k.

The centrality $C(i)$ is a measure of how close is an actor to the others and the SNA literature offers several centrality indexes [26]. The most commonly applied is the following one:

$$C(i) = \frac{G-1}{\sum_{j=1}^{G} d_{ij}} \tag{13}$$

where d_{ij} is the distance between actors i and j and G is the number of actors. $C(i)$ is the inverse of the average distance of actor i with respect to other actors and it is assumed to account for the speakers centrality.

The relative interaction of a_i with respect to another actor a_k is the following quantity:

$$f(i,k) = \frac{x_{ik} + x_{ki}}{\sum_{j=1}^{G} x_{ij} + \sum_{j=1}^{G} x_{ji}} \tag{14}$$

where the x_{ij} are the sociomatrix elements. The relative interaction accounts for the percentage of a_i interactions that have a_k as partner and, in general, $f(i,k) \neq f(k,i)$.

The AMs tend to have more interactions than the others and, for this reason, they are expected to have higher centrality. The centrality of the AM can be thought of as a random variable following a Gaussian distribution $\mathcal{N}(C_{AM}|\mu_{AM}, \sigma^2_{AM})$. If C_i is the AM centrality of sociomatrix X_i, the Maximum Likelihood (ML) estimates of μ_{AM} and σ^2_{AM} are as follows [6]:

$$\hat{\mu}_{AM} = \frac{1}{N}\sum_{i=1}^{N} C_i \quad \hat{\sigma}^2_{AM} = \frac{1}{N}\sum_{i=1}^{N} C_i^2 - \hat{\mu}^2_{AM}. \tag{15}$$

The above expressions enable one to estimate the probability of an actor a_i being the AM as follows:

$$p[r(a_i) = AM] = \mathcal{N}(C(i)|\hat{\mu}_{AM}, \hat{\sigma}^2_{AM}), \tag{16}$$

where $r(a_i)$ is the role of a_i.

The probability of a speaker a_k playing the SA role can be modeled as $p(r(a_i) = AM, r(a_k) = SA) = p(r(a_k) = SA|r(a_i) = AM)p(r(a_i) = AM)$. The second term of the product is estimated with Equation 16 and the first one can be estimated by considering that the second anchorman is the speaker that interacts more than anybody else with the AM. In other words, if a_i is the AM and a_k is the SA, the value $f(i,k)$ is expected to be, on average, higher than $f(i,j)$ for $j \neq k$. The value of $f(AM, SA)$ can be thought of as a Gaussian random variable and, if f_i is the value of $f(AM, SA)$ in sociomatrix X_i, then the ML estimates of the gaussian parameters are as follows (see above the case of the AM):

$$\hat{\mu}_{SA} = \frac{1}{N}\sum_{i=1}^{N} f_i \quad \hat{\sigma}^2_{SA} = \frac{1}{N}\sum_{i=1}^{N} f_i^2 - \hat{\mu}^2_{SA}. \tag{17}$$

The last two roles can be assigned by using the same approach as in the case of the SA. For the guests, the variable to be modeled is $f(GT, AM)$, i.e. the percentage of interactions that the guests have with the AM. For the Interview Participants, the variable is $f(IP, AM)$. The relative interactions of GT and IP with the AM are expected to follow different Gaussian distributions because GTs interact only with the AM, while IPs interact mainly with other IPs. As a result, the probability of speaker a_l being a guest can be modeled as follows:

$$\begin{gathered} p[r(a_i) = AM, r(a_k) = SA, r(a_l) = GT] = \\ p[r(a_k) = SA|r(a_i) = AM]p[r(a_i) = AM)] \\ p[r(a_l) = GT|r(a_i) = AM, r(a_k) = SA], \end{gathered} \tag{18}$$

where the last term of the product is estimated with the normal distribution: $\mathcal{N}(f(a_l, AM)|\mu_{GT}, \sigma^2_{GT})$. The expression for the probability of a speaker being an IP is the same, but GT is replaced with IP.

6 Duration Distribution Modeling

This section presents the role recognition approach based on the duration distribution analysis. Given a bulletin, each speaker k accounts for a fraction $\tau(k)$ of the total time. By the Bayes Theorem, the a-posteriori probability of a speaker k playing the role r can be written as follows:

$$p[r|\tau(k)] = \frac{p[\tau(k)|r]p(r)}{p[\tau(k)]} \tag{19}$$

where $p[\tau(k)|r]$ is the likelihood of fraction $\tau(k)$ given the role r, $p(r)$ is the a-priori probability of role r, and $p[\tau(k)]$ is the probability of observing a speaker accounting for a fraction $\tau(k)$ of a bulletin.

The approach includes some additional knowledge about the structure of the bulletins, in fact AB and MT can still be detected as the first and last speaker (see Section 5) without the need of finding the role satisfying Equation 21. Moreover, AM and SA roles can be played only by one person per bulletin, then the speaker k for which $r(k) = AM$ or $r(k) = SA$ can be indentified as follows:

$$s(r) = \arg\max_k p[\tau(k)|r]p(r) \tag{20}$$

where $s(r)$ is the speaker playing role r. The roles that can be played by different speakers in the same bulletin (GT and IP) are assigned to the remaining speakers by finding the role they identify with highest probability:

$$r(k) = \arg\max_r p[\tau(k)|r]p(r) \tag{21}$$

where $r(k)$ is the role of speaker k, and the term $p[\tau(k)]$ in Equation 19 can be neglected because it does not depend on r. This approach will be referred to as DDM in the following.

The expression of the probability distributions appearing in the right hand side of Equation 21 can be obtained as follows: $p(r)$ can be estimated with the percentage of the data in the corpus role r corresponds to. The likelihood can be estimated with a Gaussian $\mathcal{N}(\tau, \mu_r, \sigma_r)$, where averages and variances are the Maximum-Likelihood estimates obtained as follows:

$$\hat{\mu}_r = \frac{1}{N(r)} \sum_{i:r(i)=r} \tau(i) \quad \hat{\sigma}_r^2 = \frac{1}{N(r)} \sum_{i:r(i)=r} (\tau(i) - \hat{\mu}_r)^2 \tag{22}$$

where $N(r)$ is the number of speakers playing the role r in the database, and the sum involves all speakers playing role r.

The estimation of the parameters is performed using the ground truth roles provided with the recordings. In order to avoid an overestimation of the system performance, the distributions are obtained using a *leave-one-out* approach, i.e. the parameters used to process a bulletin are extracted from the whole corpus except the bulletin under examination.

7 Combination

The approaches presented so far can be used separately, but their combination is likely to improve their performance because they rely on different sources of information. In fact, SNA uses only relational data (see above) and does not take into account characteristics of single actors, while DDM uses the duration of speakers interventions and does not include any relational feature. The probability of a speaker a_i playing role r and accounting for a fraction τ of the recording time can be written as follows:

$$p[r(a_i) = r, \tau(a_i) = \tau] = p[r(a_i) = r]p[\tau(a_i) = \tau|r]p(r) \tag{23}$$

where we make the simplifying assumption that role r and fraction τ are statistically independent (see previous sections for the meaning of the symbols).

The expressions for $p[r(a_i) = r]$ are given in Section 5 for the different roles, while the expressions for $p[\tau(a_i) = \tau|r]p(r)$ are given in Section 6. The problem of role assignment can be viewed again as the maximization of vector $\vec{r}$ (see above) and this can be performed by applying the same factorization used in Section 5. The difference is that the probabilities of playing a certain role must be multiplied by the probabilities of observing a duration τ for the same speaker as shown in Equation 23.

8 Experiments and Results

This section describes the experiments performed in this work. Our approach involves two major steps (speaker clustering and role recognition) that will be evaluated with two performance measures, i.e. accuracy and purity. The next subsections describe the performance metrics, the speaker clustering results and the results of the role recognition experiments.

8.1 Performance Measures

Both speaker segmentation and role recognition give as output a sequence $S = \{(s_i, \tau_i)\}$, where $i \in (1, \dots, |S|)$, that must be compared with a groundtruth $S^* = \{(s^*_j, \tau^*_j)\}$, where $j \in (1, \dots, |S^*|)$. The number of pairs in S and S^* is not necessarily the same while the sum over the durations τ_i and τ^*_j must correspond in both cases to the total duration T of the recording from which both segmentations are extracted:

$$\sum_{i=1}^{|S|} \tau_i = \sum_{j=1}^{|S^*|} \tau^*_j = T. \tag{24}$$

In the following, s_i and s^j* will be referred to as automatic and groundtruth labels respectively, while τ_i and τ^*_j will be called automatic and groundtruth durations.

The first way to measure the segmentation performance is to consider the fraction of T such that automatic and groundtruth labels are equal. Such a measure is called *Accuracy* α and can be expressed as a percentage. If $t_i = \sum_{k=0}^{i-1} \tau_i$ is the starting point of the i^{th} automatic segment (the same applies to groundtruth segments), the duration of the intersection between segments i and j of automatic and groundtruth segmentations respectively is:

$$l_{ij} = I[min(t_{i+1}, t^*_{j+1}) - max(t_i, t^*_j)] \tag{25}$$

where $I[.]$ is equal to the argument when this is positive and zero otherwise. The accuracy can thus be calculated as follows:

$$\alpha = \frac{1}{T} \sum_{i=0}^{|S|-1} \sum_{j=0}^{|S^*|-1} l_{ij} \cdot \delta_{s_i s^*_j} \tag{26}$$

where δ_{mn} (Kronecher delta) is 1 when $m = n$ and 0 otherwise. The accuracy is a fraction and can be expressed as a percentage.

The accuracy can be applied only when the label of each segment is unique. This happens in the role segmentation because each speaker plays only one role and then has a unique label. The situation is different for the speaker segmentation because it is obtained through an unsupervised approach and the labels are simple IDs that are randomly determined and change, for the same speaker, each time the segmentation is performed. The only information in the groundtruth is then the position of the boundaries along the time axis (see Figure 2). For this reason it is necessary to use a different performance metric which is capable of measuring the consistence of the segmentation rather than the assignment of a correct label. This is the goal of the *Purity* P which measures to what extent the same speaker is labeled always with the same label and each label corresponds always to the same speaker.

Given a groundtruth speaker s^*, the purity $P(s^*)$ is defined as:

$$P(s^*) = \sum_{i=1}^{N} \frac{\sum_{k:s^*_k=s^*} l^2_{ik}}{\sum_{k:s^*_k=s^*} (\tau^*_k)^2}, \tag{27}$$

where N is the number of speakers in the automatic segmentation. The *average groundtruth purity* (π^*) is:

$$\pi^* = \sum_{s^*=1}^{N^*} \Phi(s^*) P(s^*) \tag{28}$$

where $\Phi(s^*)$ is the time fraction s^* accounts for. The same criteria can be applied to the automatic segmentation to obtain the purity of an automatic speaker s:

$$P(s) = \sum_{i=1}^{N^*} \frac{\sum_{k:s_k=s} l^2_{ik}}{\sum_{k:s_k=s} (\tau_k)^2}, \tag{29}$$

Metric	$\alpha(\%)$ bf	$\alpha(\%)$ af	$\alpha(\%)$ th	P bf	P af	P th
SNA	69.6	80.1	88.2	0.77	0.8	0.89
DDM	77.1	79.7	78.2	0.80	0.80	0.83
DDM+SNA	82.2	85.1	86.1	0.81	0.83	0.85

Table 2: Role recognition performance. This table reports the results (in terms of α and P) before filtering (bf), after filtering (af), and over the groundtruth speaker segmentation (th).

and the *average automatic purity* (π):

$$\pi = \sum_{s=1}^{N} \Phi(s)P(s). \tag{30}$$

The value of the segmentation purity P is:

$$P = \sqrt{\pi \cdot \pi^*}. \tag{31}$$

The value of P ranges between 0 and 1, but it does not correspond to a fraction, the closer to 1 the value, the more the segmentation is consistent, i.e. the same speaker tends to have the same label and, viceversa, the same label tends to be assigned to the same speaker.

8.2 Speaker Segmentation

The segmentation into speakers is the first step in the role recognition process. In fact, the sequence of the speakers is the basis for the collection of the relational data (with related sociomatrices and Social Networks) and of the segment durations.

The estimation of the parameter λ necessary to apply the smoothing algorithm is performed using a leave-one-out approach, i.e. using as a training set the whole corpus except the recording for which the parameter is estimated. In this way, no information contained in the recording is used to build the smoothing algorithm and the performance is not artificially overestimated. The use of the leave-one-out approaches enables us to use the whole corpus described in Section 3 to measure the performance of our system. The speaker clustering is not affected by the same problem because the approach we apply is unsupervised, thus there is no need to separate training and test data.

The P of the segmentation is 0.86 before the PSP based smoothing and 0.82 after. This seems to correspond to a decrease of the segmentation quality, but this is not necessarily the case when the goal is the role recognition. In fact, the average number of speaker changes in the segmentations before the PSP is 51.1, while it is 29.0 in the groundtruth. Since we use as evidence of the interaction between two speakers the fact that one of them talks immediately before the other one, such a situation introduces many spurious interactions (see central segmentation in Figure 2 and central network in Figure 4). After the PSP smoothing, the average number of speaker changes is 15.9, thus a high percentage of the groundtruth changes are lost. On the other hand, most of the preserved interactions are not spurious and this enables the system to better perform the role recognition (see lower segmentation in Figure 2 and rightmost network in Figure 4).

8.3 Role Recognition

The results of the role recognition experiments are reported in Table 2. The best performance is achieved by combining DDM and SNA and it corresponds to an accuracy of 85.1 percent and a purity 0.83 (both obtained after the PSP filtering). The performance achieved by the same system over the groundtruth speaker segmentation is $\alpha = 86.1\%$ and $P = 0.85$, then the combination captures most of the information available in the actual speaker interactions even if the speaker clustering results

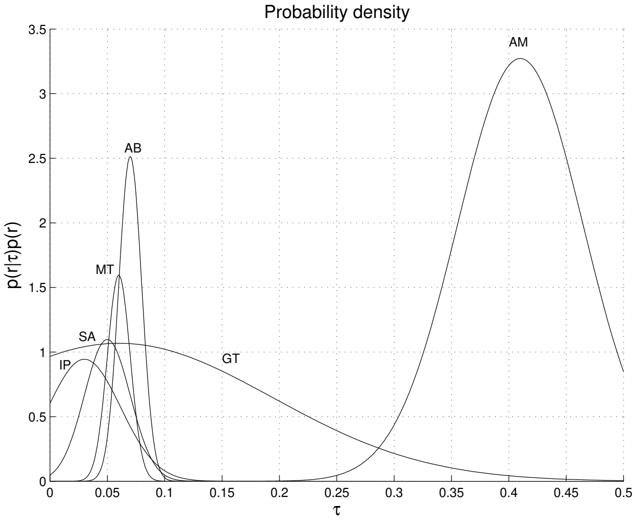

Figure 5: Probability distributions. The plot shows the τ a-posteriori probability distributions for the different roles.

in a noisy (i.e affected by errors) speaker segmentation. The filtering process improves consistently the results even if it reduces the quality of the speaker segmentation. This seems to confirm that the process preserves actual interactions while removing spurious ones. Over the groundtruth speaker segmentation, SNA performs better than DDM and DDM+SNA. The reason is that the duration models corresponding to some of the roles (see Figure 5) overlap each other and are thus ambiguous. On the other hand, DDM and SNA seem to be *diverse*, i.e. they perform different errors over the same data, then their combination improves significantly the performance of the two systems used separately.

Table 2 reports also the results obtained by applying the algorithms over the groundtruth speaker segmentation. The best results are achieved by the SNA ($\alpha = 88.2\%$), but it is the combination DDM+SNA that is closer, when applied to the automatic speaker segmentation, to the results achieved over groundtruh speaker segmentation (the difference is just one point). This seems to suggest that SNA captures better than the other algorithms the information necessary to assign roles, but it is less effective in dealing with the noise due to the speaker segmentation errors. On the contrary, the combination of SNA and DDM introduces some noise when applied to the groundtruth (the overlapping of the Gaussians in Figure 5 introduces some ambiguities), but at the same time it is more efective in dealing with the noise of the automatic speaker segmentation.

Table 3 reports the results for each role separately. Satisfactory performances are achieved for AM, GT, AB and MT, while SA and IP are recognized with low α and P. This is true especially after the PSP filtering because both SA and IP are characterized by short interventions (less than five seconds) that are often interpreted as spurious segments. This means that the problem must probably be addressed at the speaker clustering level rather than in the role recognition technique. On the other hand, the high variability of the audio data, which include backgroud noise, phone calls, crosstalk etc., makes it difficult to eliminate short spurious segments which determine the need for a filtering process that affect roles involving brief interventions. However, SA and IP account for less than ten percent of the total corpus time and the effect on the overall performance is limited.

There are several factors that can have an effect on the role recognition performance: the number G of speakers, the distance between actual speaker turns and detected speaker turns, and the oversegmentation of certain speakers, i.e. the split of a single speaker into more than one clusters.

In general, the effect of a variable x on a variable y (and viceversa) can be measured through the following *Correlation Coefficient*:

$$R = \frac{N \sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{[N \sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2][N \sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2]}} \quad (32)$$

where (x_i, y_i) are observed pairs and N is the number of pairs at disposition. When R is close to 1, the variables are correlated, i.e. when one grows the other one grows accordingly and viceversa. When R is close to 0, the two variables are not correlated and their values do not depend on each other.

The effect of G can be evaluated by estimating the correlation coefficient between G and α (in the following we use the accuracy of the combined system). In this case, $R = 0.01$, i.e. the number of actors does not affect the effectiveness of the system at least in the range of observed speakers number, i.e. from 7 to 15 (with average around 11).

The same coefficient has been used to measure the correlation between the *winDiff* of the speaker segmentation and the role recognition accuracy. The winDiff is a metric commonly used to evaluate segmentations and it is defined as follows [22]:

$$WD = \frac{1}{N(w)} \sum_{k=1}^{N(w)} |b_g(w_k) - b_a(w_k)| \quad (33)$$

where the w_k are non overlapping windows of length W that span the whole recording, $b_g(w_k)$ and $b_a(w_k)$ are the number of boundaries in groundtruth and automatic segmentation respectvely and $N(w)$ is the total number of windows covering the recording. The value of WD is the average of the difference between the number of boundaries in a given interval of automatic and groundtruth segmentations. for this reason, the lower WD the the better the automatic segmentation. The correlation coefficient between winDiff (measured with a 10 seconds long window) and accuracy is 0.18, then the distance between groundtruth speaker turns and detected speaker turns seems not to affect the performance of the system.

The situation is different when measuring the correlation between the purity for a given speaker, i.e. to what extent the speaker is always labeled with the same ID, and the purity of the corresponding role, i.e. to what extent the same speaker is given always the same, and correct, role. In this case, $R = 0.86$ and the correlation is thus significant. In other words, a speaker which is not correctly clustered is likely to be given the wrong role.

9 Conclusion

This paper has presented two approaches (applied both separately and combined) to perform role recognition in multiparty audio recordings. The experiments of this work are performed over a corpus of radio news bulletins and the results show that around 85 percent of the total time can be labeled correctly in terms of role.

Role recognition can improve several applications (the list is not exhaustive): browsing can be enhanced by enabling users to move across segments corresponding to specific roles (e.g. the meteo rather than guest interventions). In this way, it is possible to access directly information that would be difficult to locate in a long recording. Summarization systems can use the speakers role as a criterion to select the segments to be retained in an automatic summary. In the case of our data, a summary made of the only AM interventions covers all the topics presented in each bulletin and it corresponds to less than 50 percent of the recording duration. Further compression can be achieved by retaining only the AB interventions which contain a short description of the bulletin content and, on average, account for less than 10 percent of the total duration time. Role can also be used to structure audio recordings, i.e. to segment the data into intervals corresponding to sections or chapters in a text. In fact, the Social Networks correspond to the structure given to the data by the producers and can

Role	AM	SA	GT	AB	MT	IP
α SNA (bf)	94.2	30.8	49.1	33.5	98.7	47.3
α DDM (bf)	98.2	0.1	80.9	31.9	97.8	0.0
α DDM+SNA (bf)	98.2	9.6	86.9	32.2	97.8	33.0
α SNA (af)	93.2	2.9	79.6	97.6	95.4	8.9
α DDM (af)	97.3	0.0	78.6	84.3	95.4	0.0
α DDM+SNA (af)	96.2	0.0	92.7	97.8	95.4	2.5
α SNA (th)	97.9	59.1	97.8	14.5	98.4	47.8
α DDM (th)	100.0	0.0	86.7	9.0	98.4	0.0
α DDM+SNA (th)	100.0	20.0	99.9	14.6	98.4	28.2
P SNA (bf)	0.90	0.48	0.73	0.70	0.82	0.70
P DDM (bf)	0.91	0.80	0.74	0.64	0.82	0.00
P DDM+SNA (bf)	0.91	0.12	0.78	0.70	0.82	0.62
P SNA (af)	0.84	0.07	0.75	0.91	0.84	0.42
P DDA (af)	0.85	0.79	0.71	0.83	0.85	0.00
P DDM+SNA (af)	0.85	0.00	0.92	0.92	0.85	0.11
P SNA (th)	0.90	0.50	0.97	0.74	0.82	0.78
P DDA (th)	0.89	0.87	0.80	0.67	0.82	0.00
P DDM+SNA (th)	0.90	0.17	0.86	0.74	0.82	0.58

Table 3: Performance per role. The results are reported both before the filtering(bf), after the filtering (af), and using the groundtruth (th).

be used to detect the structure a-posteriori. Moreover, any application involving an indexing step (Information Retrieval, Categorization, Topic detection, Thematic Indexing, etc.) can use the role as an index or as an information related to the audio content.

The approaches presented in this work are especially suitable for data created following a plan or a structure which remains stable across different recordings. This is the case for most radio or television programs which are produced using a *format* assigning specific roles and tasks to every person. Moreover, the role recognition is useful only when the number of individuals is high enough, i.e. more than 8-10 persons, to build structures that can be detected through SNA and duration distributions. In the case of data that are not the result of a plan (e.g. home-made videos), the same approaches can still help to cluster recordings with similar structure or to detect individuals that are more central than others.

The main limit of the approach we propose is that it can be used only for data involving the same roles and the same kind of interaction patterns, i.e. a central individual dealing directly with most of the other persons. On the other hand, our work shows that interaction patterns carry information that so far, to our knowledge, has been neglected and can be used to better represent the content of the data. The measures introduced in Section 5 (centrality and interaction fraction) seem to account for important aspects of people activity and can probably be used in other contexts to recognize other roles (e.g. the chairman in meetings or conference calls), collective actions (e.g. discussions or question answerings in lectures), etc.

In the case of our data, the SNs have a compact structure, i.e. there is a single central individual with a more or less direct relationship with all other persons. However, other programs have a more complex structure where it is possible to identify several central figures playing a major role only in part of the recording. This requires to include in the network analysis tasks like the detection of *cohesive subgroups* and *dyadic or triadic models* [26] that will be the subject of future work.

References

[1] J. Ajmera. *Robust Audio Segmentation.* PhD thesis, École Polytechnique Fédérale de Lausanne (EPFL), 2004.

[2] J. Ajmera and C. Wooters. A robust speaker clustering algorithm. In *Proceedings of IEEE Workshop on Automatic Speech Recognition Understanding*, 2003.

[3] N. Babaguchi, Y. Kawai, and T. Kitahashi. Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE Transactions on Multimedia*, 4(1):68–75, 2002.

[4] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison Wesley, 1999.

[5] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker. The rules behind the roles: identifying speaker roles in radio broadcasts. In *Proceedings of American Association of Artificial Intelligence Symposium*, 2000.

[6] C.M. Bishop. *Pattern Recognition and Machine Learning.* Springer Verlag, 2006.

[7] C.H. Chan and G.J.F. Jones. Affect-based indexing and retrieval of films. In *Proceedings of ACM Conference on Multimedia*, pages 427–430, 2005.

[8] F. Coldefy and P. Bouthemy. Unsupervised soccer video abstraction based on pitch, dominant color and camera motion analysis. In *Proceedings of ACM Conference on Multimedia*, pages 268–271, 2004.

[9] N. Fay, S. Garrod, and J. Carletta. Group discussion as interactive dialogue or as serial monologue. *Psychological Science*, 11(6):481–486, 2000.

[10] D. Gatica-Perez, D. Zhang, and S. Bengio. Extracting information from multimedia meeting collections. In *Proceedings of 7^{th} ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 242–252, 2005.

[11] A. Hanjalic. Adaptive extraction of highlights from a sport video based on excitement modeling. *IEEE Transactions on Multimedia*, 7(6):1114–1122, 2005.

[12] A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.

[13] X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing: A guide to theory, algorithm and system development.* Prentice Hall, 2001.

[14] D.O. Javed, Z. Rasheed, and M. Shah. A framework for segmentation of talk and game shows. In *Proceedings of International Conference on Computer Vision*, 2001.

[15] F. Jelinek. *Statistical Methods for Speech Recognition.* MIT Press, 1997.

[16] K. Koumpis and S. Renals. Content-based access to spoken audio. *IEEE Signal Processing Magazine*, 22(5):61–69, 2005.

[17] C.M. Lee and S.S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Multimedia*, 13(2):293–303, 2005.

[18] S.L. Lee and B. Chen. Spoken document understanding and organization. *IEEE Signal Processing Magazine*, 22(5):42–60, 2005.

[19] Y.-F. Ma, X.-S Hua, L. Lu, and H.-J. Zhang. A generic framework for user attention model and its application in video summarization. *IEEE Transactions on Multimedia*, 7(5):907–919, 2005.

[20] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):305–317, 2005.

[21] A. Papoulis. *Probability, Random Variables, ans Stochastic Processes*. McGraw Hill, 1991.

[22] L. Pevzner and M.A. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, 2002.

[23] J. Scott. *Social Network Analysis - A handbook*. SAGE Publications, 1991.

[24] C.G.M. Snoek and M. Worring. Multimedia event-based video indexing using time intervals. *IEEE Transactions on Multimedia*, 7(4):638–647, 2005.

[25] C.G.M. Snoek and M. Worring. Multimodal video indexing: a review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, 2005.

[26] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, 1994.

[27] S.N. Wrigley, G.J. Brown, V. Wan, and S. Renals. Speech and crosstalk detection in multichannel audio. *IEEE Transactions on Speech and Audio Processing*, 13(1):84–91, 2005.

[28] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan. Modeling individual and group actions in meetings with layered HMMs. *accepted for publication by IEEE Transactions on Multimedia, to appear*, 2006.

[29] T. Zhang and C.C. Jay Kuo. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on Multimedia*, 9(4):441–457, 2001.