Automatic Role Recognition in Multiparty Recordings: Using Social Affiliation Networks for Feature Extraction

Hugues Salamin, Sarah Favre, *Member, IEEE*, and Alessandro Vinciarelli

***Abstract*—Automatic analysis of social interactions attracts increasing attention in the multimedia community. This letter considers one of the most important aspects of the problem, namely the roles played by individuals interacting in different settings. In particular, this work proposes an automatic approach for the recognition of roles in both production environment contexts (e.g., news and talk-shows) and spontaneous situations (e.g., meetings). The experiments are performed over roughly 90 h of material (one of the largest databases used for role recognition in the literature) and show that the recognition effectiveness depends on how much the roles influence the behavior of people. Furthermore, this work proposes the first approach for modeling mutual dependences between roles and assesses its effect on role recognition performance.**

***Index Terms*—Broadcast data, meeting recordings, role recognition, social network analysis.**

I. Introduction

THE computing community is making significant efforts towards the development of automatic approaches for the analysis of social interactions (see [1]–[3] for extensive surveys of the domain). This is not surprising as social interactions are not only one of the most important aspects of our everyday lives but also a ubiquitous subject in multimedia data: radio and television programs (debates, news, talk-shows, movies, etc.) rarely show something other than social interactions. The way people interact depends on the context, but there is one aspect that all social interactions seem to have in common:

People do not interact with one another as anonymous beings. They come together in the context of specific environments and with specific purposes. Their interactions involve behaviors associated with defined statuses and particular roles. These statuses and roles help to pattern our social interactions and provide predictability [4].

As the above suggests that roles are a universal key to understand social interactions and these are one of the most common subjects of multimedia material, this work proposes an approach for the automatic recognition of roles in multiparty recordings.

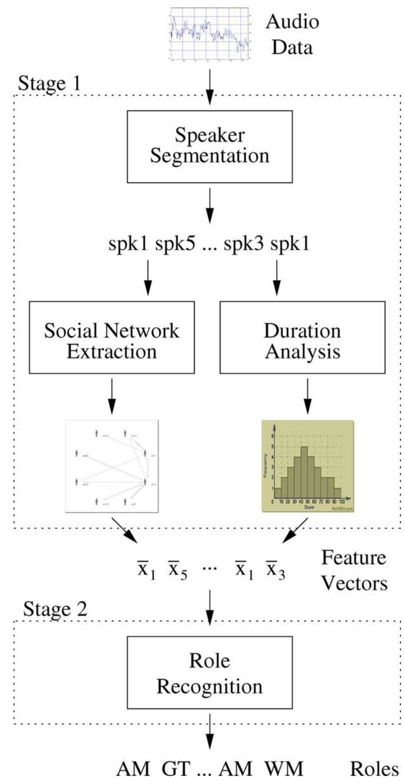

Fig. 1. Role recognition approach. The picture shows the two main stages of the approach: the features extraction and the actual role recognition.

The approach includes two main stages (see Fig. 1): the first is the *feature extraction* and it involves the automatic construction of a Social Affiliation Network (SAN)[5] as well as its conversion into feature vectors that represent each person in terms of their relationships with the others. The second stage is the *role recognition*, i.e., the mapping of the feature vectors extracted in the first stage into roles belonging to a predefined set. This task is performed using Bernoulli or Multinomial distributions [6] for the Affiliation Network features and Gaussian distributions for the intervention lengths associated to each role.

The experiments have been performed over three different corpora (see Section V-A for more details): a collection of radio news bulletins (around 20 h), a dataset of radio talk-shows (around 25 h), and the AMI meeting corpus (around 45 h) [7]. To the best of our knowledge, there is only one work reporting experiments performed over a larger amount of data [8]. However, the corpus of [8] includes only the news scenario, while our data include other settings. This is important because it allows one to assess the approach robustness with respect to changes of the interaction structure.

Manuscript received June 08, 2009; revised July 19, 2009. First published August 21, 2009; current version published October 16, 2009. This work was supported in part by the Swiss National Science Foundation (under the National Centre of Competence in Research on Interactive Multimodal Information Management), and in part by the European Community's Seventh Framework Programme (FP7/2007-2013), under grant agreement no. 231287 (SSPNet). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xian-Sheng Hua.

H. Salamin and A. Vinciarelli are with the Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, UK. When the work was done, H. Salamin and A. Vinciarelli were with Idiap Research Institute (e-mail: hugues.salamin@dcs.gla.ac.uk; alessandro.vinciarelli@dcs.gla.ac.uk).

S. Favre is with Idiap Research Institute, CP592 1920 Martigny, Switzerland, and also with École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland (e-mail: sfavre@idiap.ch).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2009.2030740

For the first two datasets, the accuracy (percentage of recording time correctly labeled in terms of roles) ranges from 60% to 85%; for the third dataset, the accuracy is around 45%. One possible explanation of the difference is that roles are easier to model when they are *formal*, i.e., correspond to functions that impose more or less rigorous constraints on the way people behave and interact with the others (like in the case of broadcast data). In contrast, roles are harder to model when they are *informal*, i.e., when they correspond to a position in a given social system (e.g., manager in a company) and do not necessarily impose tight constraints on the way people behave and interact (like in the case of meetings). However, the performance significantly outperforms chance for both broadcast and meeting recordings.

Role recognition can be useful in several applications (the list is not exhaustive). For example in media browsers, the information about the role of the person speaking at a given moment can help users to quickly identify segments of interest. In summarization, the role of people can be used as a criterion to select representative segments of the data [9], [10]. In information retrieval, the role can be used as an index to enrich the content description of the data. Furthermore, the role can be used to segment the data into semantically coherent segments [11], [12].

The main contributions of this paper with respect to previous approaches proposed by the authors [13] and the rest of the literature are as follows.

- The approach proposed in [13] can be applied only to groups involving at least eight to ten persons because it is based on simple Social Networks and these need at least this number of people to produce meaningful features. This work addresses such a limit by introducing the use of SANs, a different kind of network that makes it possible to analyze smaller groups. Without this change, the analysis of the AMI meetings (including only four participants) would not be possible.
- The approach in [13] does not take into account the dependence between roles. Each person is assigned a role independently of those assigned to others. This work proposes an approach to overcome this limit and takes into account the constraints that the role distribution across different interacting participants must respect. To the best of our knowledge, this is a novelty not only with respect to [13] but also with respect to the state-of-the-art.
- To the best of our knowledge, this is the first work in the literature that reports experiments performed over different interaction contexts, i.e., production environment data involving formal roles (news and talk-shows) and spontaneous settings involving informal roles (meetings).

The rest of the paper is organized as follows: Section II presents a survey of related works, Section III describes the feature extraction stage, Section IV describes the role recognition stage, Section V presents experiments and results, and Section VI draws some conclusions.

II. Related Work

Role recognition works presented in the literature (see [1] and [3] for survey) can be split into two major groups depending on whether they address the recognition of *formal* or *informal* roles [14]. The former corresponds to specific functions to be fulfilled in a given social context (e.g., the *chairman* in a meeting) and tends to induce stable, machine detectable, behavioral patterns. The latter corresponds to positions in a social system (e.g., the *manager* in a company) and does not necessarily result into detectable behavioral patterns. Table I reports claimed performance and basic data descriptions for each work discussed in this section.

Most of the works dedicated to formal roles perform experiments over *production environment* data like movies, news, talk-shows, etc. Some approaches [8], [15] apply techniques like hidden Markov models or boosting and use features accounting for the speaking activity of people, e.g., intervention length, number of interventions, lexical choices (distributions of bigrams and trigrams), etc. Other approaches [13], [16] have proposed the use of Social Networks as a mean to extract features that are given as input to Bayesian classifiers [13] or used to build co-occurrence matrices aimed at identifying social groups [16].

The recognition of informal roles is typically performed using meeting recordings. The work in [17] recognizes social roles suggested by human sciences (e.g., *gate-keeper* or *attacker*) by feeding support vector machines with features extracted from both audio and video. These include the same features described above for formal roles and *fidgeting* measures extracted from the video. The approaches in [18] and [19] are tested over the same meeting data as those used in this work (see Section V-A). The first work combines a Bayesian classifier fed with features extracted using Social Networks, and boosting techniques applied to the distribution of words, bigrams, and trigrams extracted from the automatic transcriptions of the interventions. The second work uses speaking activity features (e.g., probability of initiating a talk-spurt when someone else is speaking or when a participant in a specific other role is speaking). The AMI meeting corpus has been used as well for automatic recognition of dominant clique (the two most dominant persons) [20] and relationship between dominance and one of the roles played in the corpus (the *Project Manager*) [21]. While these two works cannot be said to address specifically the role recognition problem, they still are similar to the others presented in this section as they identify persons with specific social characteristics depending on their behavior.

III. Feature Extraction

This section presents the feature extraction stage aimed at extracting and representing the interaction pattern of each person. The stage includes two steps: the first is the segmentation of the recordings into single speaker segments (speaker diarization), and the second is the extraction of a SAN from the resulting speaker sequence (see upper dotted box in Fig. 1).

The experiments involve two kinds of data: radio programs, where there is a single audio channel, and meeting recordings, where each participant wears a headset microphone. This requires the application of different speaker diarization techniques fully described in [22] (broadcast data) and [23] (meeting recordings). The techniques are not described here because they are not the main element of interest in this work. Section III-A shows how the output of the speaker diarization

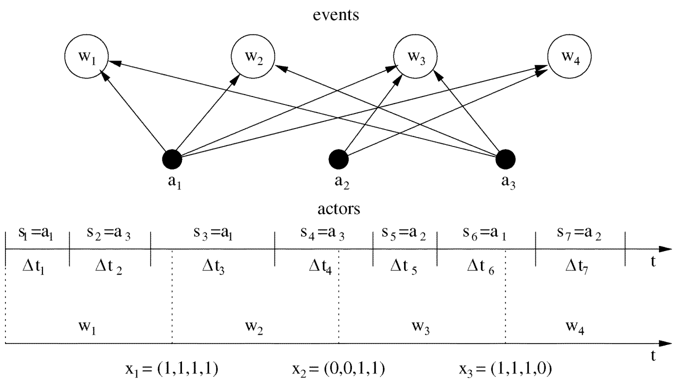

Fig. 2. Interaction pattern extraction. The picture shows the SAN extracted from a speaker segmentation. The events of the network correspond to the segments w_j and the actors are linked to the events when they talk during the corresponding segment. The actors are represented using n-tuples $\vec{x}_a$ where the components account for the links between actors and events.

is used to build a SAN and represent people with n-tuples accounting for their interaction pattern.

A. *Affiliation Network Extraction*

The result of the speaker diarization process is that each recording is split into a sequence $S = \{(s_i, \Delta t_i)\}$, where $i \in \{1, \ldots, |S|\}$, s_i is the label assigned to the speaker voice detected in the ith segment of audio, and Δt_i is the duration of the ith segment. The label s_i belongs to the set A of unique speaker labels, outputed by the speaker diarization process (see lower part of Fig. 2). The sequences extracted from the speaker diarization are used to create a SAN representing the relationships between the roles. A SAN is a graph with two kinds of nodes: the *actors* and the *events* [5]. Actors can be linked to events, but no links are allowed between nodes of the same kind (see upper part of Fig. 2). In the experiments, the actors correspond to the people involved in the recordings, and the events correspond to uniform non-overlapping segments spanning the whole length of the recordings. The rationale behind this choice is that actors speaking in the same interval of time are more likely to talk with one another (i.e., of interacting with one another) than actors speaking in different intervals of time. Thus, the SAN encodes information about *who interacts with whom and when*.

One of the main advantages of this representation is that each actor a can be represented by a n-tuple $\mathbf{x}_a = (x_{a1}, \ldots, x_{aD})$, where D is the number of segments used as events and the component x_{aj} accounts for the participation of the actor a in the jth event. The experiments make use of two kinds of representation. In the first one, component x_{aj} is 1 if the actor a talks during the jth segment and 0 otherwise (the corresponding n-tuples are shown at the bottom of Fig. 2). In the second one, x_{aj} is the number of times that actor a talks during the jth segment. In the first case, the n-tuples are binary, and in the second case, they have integer components higher or equal to 0. In both cases, people that interact more with each other tend to talk during the same segments and are represented by similar n-tuples. If the roles influence the structure of the relationships between people, similar n-tuples should correspond to the same role.

IV. Role Recognition

The problem of role recognition can be formalized as follows: given a set of actors A and a set of roles $\mathcal{R}$, find the function $\varphi : A \rightarrow \mathcal{R}$ mapping the actors into their actual role. In other words, the problem corresponds to finding the function φ such that $\varphi(a)$ is the role of actor a.

Section III has shown that the interaction pattern of each actor a is represented with a n-tuple $\mathbf{x}_a = (x_{a1}, \ldots, x_{aD})$, where D is the number of segments, that can have either binary or positive integer components. Furthermore, every actor a talks for a fraction τ_a of the total time of the recording. Thus, each actor corresponds to a pair $\mathbf{y}_a = (\tau_a, \mathbf{x}_a)$.

Given a function $\varphi : A \rightarrow \mathcal{R}$ and the set of observations $Y = \{\mathbf{y}_a\}_{a \in A}$, the problem of assigning a role to each actor can be thought of as the maximization of the *a-posteriori* probability $p(\varphi|Y)$. By applying the Bayes Theorem and by taking into account that $p(Y)$ is constant during recognition, this problem is equivalent to finding $\hat{\varphi}$ such that

$$\hat{\varphi} = \arg\max_{\varphi \in \mathcal{R}^A} p(Y|\varphi)p(\varphi) \tag{1}$$

where $\mathcal{R}^A$ is the set of all possible functions mapping actors into roles.

In order to simplify the problem, two assumptions are made: the first is that the observations are mutually conditionally independent given the roles. The second is that the observation $\mathbf{y}_a$ of actor a only depends on its role $\varphi(a)$ and not on the role of the other actors. Equation (1) can thus be rewritten as

$$\hat{\varphi} = \arg\max_{\varphi \in \mathcal{R}^A} p(\varphi) \prod_{a \in A} p(\mathbf{y}_a|\varphi(a)). \tag{2}$$

The above expression is further simplified by assuming that the speaking time τ_a and the interaction n-tuples $\mathbf{x}_a$ of actors a are statistically independent given the role $\varphi(a)$; thus, the last equation becomes

$$\hat{\varphi} = \arg\max_{\varphi \in \mathcal{R}^A} p(\varphi) \prod_{a \in A} p\left(\mathbf{x}_a|\varphi(a)\right) p(\tau_a|\varphi(a)). \tag{3}$$

The probabilities appearing in the last equation have been estimated using different models to take into account the two representations of $\mathbf{x}_a$ described above, and to model the constraints in the distribution of roles (e.g., there must be only one *anchorman* in a given talk-show), i.e., to explicitly take into account the dependence between the roles.

The next sections show how $p(\mathbf{x}_a|\varphi(a))$, $p(\tau_a|\varphi(a))$, and $p(\varphi)$ are estimated in the experiments.

A. *Modeling Interaction Patterns*

This section shows how the probability $p(\mathbf{x}_a|\varphi(a))$ is estimated for both binary and multinomial n-tuples $\mathbf{x}_a$ (see Section III-A).

When the components of the n-tuple $\mathbf{x}_a$ are binary, i.e., $x_{aj} = 1$ when actor a talks during segment j and 0 otherwise, the most natural way of modeling $\mathbf{x}_a$ is to use independent Bernoulli discrete distributions:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{j=1}^{D} \mu_j^{x_j}(1-\mu_j)^{1-x_j} \tag{4}$$

where D is the number of events in the network (see Section III), and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_D)$ is the parameter vector of the distribution. A different Bernoulli distribution is trained for each role. The maximum likelihood estimates of the parameters $\boldsymbol{\mu}_r$ for a given role r are as follows [6]:

$$\mu_{rj} = \frac{1}{|A_r|}\sum_{a\in A_r} x_{aj} \tag{5}$$

where A_r is the set of actors playing the role r in the training set, and $\mathbf{x}_a$ is the n-tuple representing the actor a.

When the components $\mathbf{x}_j$ correspond to the number of times that actor a talks during event j, i.e., when the components are integers greater or equal to 0, they can be represented with a vector $\mathbf{z}_i = (z_{i1}, \ldots, z_{iT})$, where T is the maximum number of times that an actor can talk during a given event, $z_{ij} \in \{0, 1\}$, and $\sum_{j=1}^{T} z_{ij} = 1$ (*one-out-of-*K). In other words, x_i is represented with a T-dimensional vector where all the components are 0 except one, i.e., the component $z_{in} = 1$, where n is the number of times that the actor represented by $\mathbf{x}$ talks during event i. As a result, $\mathbf{x}$ is represented as a n-tuple of vectors $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_D)$ and can be modeled as a product of independent multinomial distributions:

$$p(\mathbf{z}|\boldsymbol{\mu}) = \prod_{i=1}^{D}\prod_{j=1}^{T} \mu_{ij}^{z_{ij}}. \tag{6}$$

The parameters $\boldsymbol{\mu}$ can be estimated by maximizing the likelihood of $p(\mathbf{z}|\boldsymbol{\mu})$ over a training set $\mathcal{X}$. This leads to a closed form expression for the parameters:

$$\mu_{ij} = \frac{1}{|A_r|}\sum_{a\in A_r} z_{aij}. \tag{7}$$

B. Modeling Durations

Given a labeled training set, there is a set A_r of actors playing role r, $p(\tau|r)$ is estimated using a Gaussian distribution $\mathcal{N}(\tau|\mu_r, \sigma_r)$, where μ_r and σ_r are the sample mean and variance, respectively:

$$\mu_r = \frac{1}{|A_r|}\sum_{a\in A_r} \tau_a \tag{8}$$

$$\sigma_r = \frac{1}{|A_r|}\sum_{a\in A_r} (\tau_a - \mu_r)^2. \tag{9}$$

This corresponds to a maximum likelihood estimate, where a different Gaussian distribution is obtained for each role.

C. Estimating Role Probabilities

This subsection shows how the *a-priori* probability $p(\varphi(a))$ of actor a playing role $\varphi(a)$ is estimated. Two approaches are proposed: the first is based on the assumption that roles are independent and does not take into account the constraints that the role distribution across different participants in a given recording must respect, e.g., there is only one *Anchorman* in a talk-show, there is only one *Project Manager* in a meeting, etc. The second approach considers the roles to be dependent and takes into account the above constraints.

1) Modeling Independent Roles: The first approach assumes that the roles are independent and thus that $p(\varphi)$ is simply the product of the *a-priori* probabilities of the roles assigned through φ to the different actors:

$$p(\varphi) = \prod_{a\in A} p(\varphi(a)). \tag{10}$$

The *a-priori* probability of observing the role r can be estimated as follows:

$$p(\varphi(a)) = \frac{N_{\varphi(a)}}{N} \tag{11}$$

where N and $N_{\varphi(a)}$ are the total number of actors and the total number of actors playing role $\varphi(a)$ in the training set.

Using the above approach, Equation (2) boils down to

$$\hat{\varphi} = \arg\max_{\varphi\in\mathcal{R}^A} \prod_{a\in A} p(\mathbf{x}_a|\varphi(a))p(\tau_a|\varphi(a))p(\varphi(a)) \tag{12}$$

and the role recognition process simply consists in assigning each actor the role $\varphi(a)$ that maximizes the probability $p(\mathbf{x}_a|\varphi(a))p(\tau_a|\varphi(a))p(\varphi(a))$.

2) Modeling Dependent Roles: The second approach tries to model the constraints that the role distribution of a given recording must respect. For example, there must be only one *Anchorman* in a talk-show while the number of *Guests* can change at each edition of the talk-show. In this case, the roles played by the different recording participants cannot be considered independent, and $p(\varphi)$ cannot be written as the product of the *a-priori* probabilities of the roles [like in (10)].

A given mapping $\varphi \in \mathcal{R}^A$ corresponds to a distribution of roles across the different recording participants where each role is played by a certain number of actors. The constraints to be respected are expressed in terms of the number of actors that can play a given role (e.g., only one actor can be the *Anchorman*). Thus, $p(\varphi)$ must be different from 0 only for those distributions of roles that respect the constraints. The number of possible actors playing some roles is actually predetermined (i.e., exactly n_r actors must play role r), while for others, the only available *a-priori* information is that at least one person must play the role (i.e., $n_r > 0$).

TABLE I
SYNOPSIS OF ROLE RECOGNITION RESULTS. THE TABLE PROVIDES A BRIEF DESCRIPTION OF THE DATA USED IN THE LITERATURE, AS WELL AS THE PERFORMANCE ACHIEVED IN THE DIFFERENT WORKS

Ref.	Data	Time	Roles	Performance
[8]	TDT4 Mandarin broadcast news (336 shows, 3 roles)	170h.00m	formal	77.0% of the news stories correctly labeled in terms of role
[13]	Radio news bulletins (96 recordings, 6 roles)	25h.00m	formal	85% of the data time correctly labeled in terms of role
[15]	NIST TREC SDR Corpus (35 recordings, publicly available 3 roles)	17h.00m	formal	80.0% of the news stories correctly labeled in terms of role
[16]	Movies and TV shows (10 movies and 3 TV shows , 9-20 roles)	21h.00m	formal	95% of leading roles correctly assigned and 84.3% of community roles correctly assigned
[17]	The Mission Survival Corpus (11 recordings, publicly available, 5 roles)	4h.30m	informal	90% of analysis windows (around 10 seconds long) correctly classified in terms of task area roles and 95% in terms of socio area roles
[18]	AMI Meeting Corpus (138 recordings, publicly available, 4 roles)	45h.00m	informal	53% of the data time correctly labeled in terms of role
[19]	AMI Meeting Corpus (138 recordings, publicly available, 4 roles)	45h.00m	informal	67.9% of the data time correctly labeled in terms of role

TABLE II
ROLE DISTRIBUTION. THE TABLE REPORTS THE PERCENTAGE OF TIME EACH ROLE ACCOUNTS FOR IN C1, C2, AND C3

Corpus	AM	SA	GT	IP	HR	WM	PM	ME	UI	ID
C1	41.2%	5.5%	34.8%	4.0%	7.1%	6.3%	N/A	N/A	N/A	N/A
C2	17.3%	10.3%	64.9%	0.0%	4.0%	1.7%	N/A	N/A	N/A	N/A
C3	N/A	N/A	N/A	N/A	N/A	N/A	36.6%	22.1%	19.8%	21.5%

According to the above, $p(\varphi)$ is modeled with a product of multinomial distributions [6]:

$$p(\varphi) = \prod_{r \in \mathcal{R}} p(\mathbf{z}_r | \boldsymbol{\mu}_r) \tag{13}$$

where $\mathbf{z}_r$ is a *one-out-of-K* (see Section IV-A) representation of the number of times a role can be played in a given recording, and $\boldsymbol{\mu}_r$ is the parameter vector.

We can divide the set $\mathcal{R}^A$ in classes $\{C_g\}$ where all mappings lead to a role distribution where the same role is played always the same number of times. We assume that all mappings φ in the same class have the same probability. Thus, the probability of observing a given assignment is

$$p(\varphi) = \frac{\prod_{r \in \mathcal{R}} p(\mathbf{z}_r | \boldsymbol{\mu}_r)}{|C_g|}. \tag{14}$$

Then in the second model, (2) can be rewritten as

$$\hat{\varphi} = \arg \max_{\varphi \in \mathcal{R}^A} p(\varphi) \prod_{a \in A} p(\mathbf{x}_a | \varphi(a)) p(\tau_a | \varphi(a)) \tag{15}$$

where $p(\varphi)$ is the expression of (14). Maximizing this product using a brute-force approach is not tractable if the number of actors is high. Therefore, we used simulated annealing [24] to approximate the best mapping for each recording.

V. EXPERIMENTS AND RESULTS

The next four sections describe data and roles, performance measures, experimental setup, and role recognition results.

A. Data and Roles

The experiments of this work have been performed over three different corpora referred to as C1, C2, and C3 in the following. C1 contains all news bulletins (96 in total) broadcasted by *Radio Suisse Romande* (the French-speaking Swiss national broadcasting service) during February 2005. The average length of C1 recordings is 11 min and 50 s, and the average number of participants is 12. C2 contains all talk-shows (27 in total) broadcasted by *Radio Suisse Romande* during February 2005. All C2 recordings are one hour long and the average number of participants is 25. C3 is the AMI meeting corpus [7], a collection of 138 meeting recordings involving four persons each and with an average length of 19 min and 50 s. While C1 and C2 contain real-world news and talk-shows, the meetings in C3 are a *simulation* and the participants act roles they do not play in their real life.

The roles of C1 and C2 share the same names and correspond to similar functions: the *Anchorman* (AM), i.e., the person managing the program, the *Second Anchorman* (SA), i.e., the person supporting the AM, the *Guest* (GT), i.e., the person invited to report about a single and specific issue, the *Interview Participant* (IP), i.e., interviewees and interviewers, the *Headline Reader* (HR), i.e., the speaker reading a short abstract at the beginning of the program, and the *Weather Man* (WM), i.e., the person reading the weather forecasts. However, even if the roles have the same name and correspond to roughly the same functions, they are played in a different way in C1 and C2 (e.g., consider how different is the behavior of an anchorman in news supposed to inform and in talk-shows supposed to entertain). In C3, the role set is different and contains the *Project Manager* (PM), the *Marketing Expert* (ME), the *User Interface Expert* (UI), and the *Industrial Designer* (ID). See Table II for the distribution of roles in the corpora.

B. Speaker Diarization Results

The interaction patterns used at the role recognition step are extracted from the speaker segmentation obtained with the two different diarization processes (see Section III). Errors in the diarization (e.g., people detected as speaking when they are silent, or multiple voices attributed to a single speaker) lead to spurious interactions that can mislead the role recognition process.

The effectiveness of the diarization is measured with the *purity* π, a metric showing on one hand to what extent all feature vectors corresponding to a given speaker are detected as belonging to the same voice, and on the other hand to what extent all vectors detected as a single voice actually correspond to a single speaker. The purity ranges between 0 and 1 (the higher the better) and it is the geometric mean of two terms: the *average cluster purity* π_c and the *average speaker purity* π_s. The definition of π_c is as follows:

$$\pi_c = \sum_{k=1}^{N_c} \sum_{l=1}^{N_s} \frac{n_k}{N} \frac{n_{lk}^2}{n_k^2} \tag{16}$$

where N is the total number of feature vectors, N_s is the number of speakers, N_c is the number of voices detected in the diarization process, n_{lk} is the number of vectors belonging to speaker l that have been attributed to voice k, and n_k is the number of feature vectors in voice k. The definition of π_s is as follows:

$$\pi_s = \sum_{l=1}^{N_s} \sum_{k=1}^{N_c} \frac{n_l}{N} \frac{n_{lk}^2}{n_l^2} \tag{17}$$

(see above for the meaning of the symbols).

The average purity is 0.81 for C1 and 0.79 for C2. The average purity for C3 is 0.99. The difference in purity is explained by the different experimental conditions and methods used to obtain the speaker segmentation.

C. Experimental Setup

The experiments are based on a K-fold cross-validation approach [6]. The corpora are split into K equally sized parts of which $K-1$ are used as training set, while the remaining one is used as the test set. Each of the K parts is used iteratively as the test set so that the experiments can be performed over the whole dataset while still preserving a rigorous separation between training and test set. In the case of our experiments, $K = 5$ and each subset contains 20% of the data. The only hyperparameter to be set is the number D of segments used as events in the SAN. At each iteration of the K-fold cross-validation, D is varied such that the value giving the highest role recognition results *over the training set* has been retained for testing. *In this way, a rigorous separation between the training and test set has been observed for the setting of the hyperparameter as well.*

The statistical significance of performance differences is assessed with the Kolmogorov–Smirnov test [25]. The advantage of this test is that it does not make assumptions about the distribution of the performance (unlike the t-test that assumes the performances following a Gaussian distribution) and it is adapted to continuous distributions (unlike the χ^2-test that requires the distributions to be made discrete through histogramming).

D. Role Recognition Results

Table III reports the results achieved over C1 and C2, Table IV those obtained for C3. The performance is measured in terms of *accuracy*, i.e., the percentage of time correctly labeled in terms of role in the test set. Each accuracy value is accompanied by the standard deviation of the accuracies achieved over the different recordings of each corpus. The distribution used to model the interaction patterns is indicated with *B* (Bernoulli) and *M* (multinomial). The approach used to estimate the *a-priori* role probabilities is indicated with *I* (independence) and *D* (dependence).

Modeling the dependence between roles leads to statistically significant improvements for C2 and C3, while it decreases the performance for C1. One probable explanation is that C1 presents more variability in the number of people playing a given role; thus, $p(\varphi)$ (see Section IV-C) cannot be estimated as reliably as for the other corpora. However, these results suggest that taking into account the dependence across roles is beneficial as long as $p(\varphi)$ can be estimated reliably. To the best of our knowledge, this is the first attempt to model explicitly the dependence between roles and the results provide a first assessment of what can be expected, at least for the approach proposed here, in terms of performance improvement.

For the three corpora, *the differences between the performances achieved using Bernoulli and multinomial distributions are not statistically significant*. This suggests that the important information is presence/absence (conveyed by the Bernoulli distribution) and not number of times a speaker talks during an event (conveyed by the multinomial). This is not surprising because the most important aspect encoded by SANs (at least for the approach proposed in this work) is who interacts with whom and not how much someone interacts with someone else.

Overall, roles in meeting data appear to be harder to model for several reasons. On one hand, roles in meetings are *informal*, i.e., they correspond to a position in a given social system and do not correspond to stable behavioral patterns like in the case of the *formal* roles in broadcast data. On the other hand, the meetings in C3 are not real-world, i.e., the participants *act* in a scenario that does not correspond to their real lives. Not surprisingly, the meeting role recognized with highest accuracy is the *Project Manager* (PM). In fact, the PM plays also the role of *chairman*, i.e., a formal role that influences the actual interaction pattern of the people that play it. *The performance difference when passing from manual (ground truth) to automatic speaker diarization is statistically significant for C1 and C2* (see Tables III and IV). The difference is not significant for C3 because the purity of the speaker segmentation for such a corpus is 0.99, i.e., it corresponds almost perfectly to the groundtruth speaker segmentation. In contrast, the difference is significant for C1 and C2 because in this case, the speaker diarization process produces more errors and the purity is around 0.8, i.e., the output of the speaker diarization is significantly different from the groundtruth speaker segmentation. The difference in accuracy is around 10% (statistically significant) and this is mostly due to the small differences (2 s on average)

TABLE III
ROLE RECOGNITION PERFORMANCE FOR C1 AND C2. THE TABLE REPORTS BOTH THE OVERALL ACCURACY AND THE ACCURACY FOR EACH ROLE. "B" STANDS FOR *BERNOULLI*, "M" STANDS FOR *MULTINOMIAL*, "I" STANDS FOR ROLES *INDEPENDENCE*, AND "D" STANDS FOR ROLES *DEPENDENCE*. THE OVERALL ACCURACY IS ACCOMPANIED BY THE STANDARD DEVIATION σ OF THE PERFORMANCES ACHIEVED OVER THE SINGLE RECORDINGS. THE UPPER PART OF THE TABLE REPORTS THE RESULTS OBTAINED OVER THE OUTPUT OF THE SPEAKER SEGMENTATION, THE LOWER PART REPORTS THE RESULTS OBTAINED OVER THE MANUAL SPEAKER SEGMENTATION

	all (σ)	AM	SA	GT	IP	HR	WM
Automatic Speaker Segmentation							
C1 (B,I)	81.7 (6.9)	98.0	4.0	92.0	5.6	55.9	76.8
C1 (B,D)	62.7 (16.5)	89.9	4.2	68.9	9.0	11.0	10.1
C1 (M,I)	**82.4** (7.1)	97.8	4.8	92.2	4.2	64.3	78.2
C1 (M,D)	62.3 (16.7)	88.7	3.4	70.2	4.5	7.0	15.4
C2 (B,I)	83.2 (6.7)	75.0	88.3	91.5	N/A	29.1	9.0
C2 (B,D)	87.5 (4.4)	77.1	92.1	93.2	N/A	91.0	17.7
C2 (M,I)	84.0 (6.5)	68.7	92.2	89.7	N/A	83.7	15.4
C2 (M,D)	**87.8** (4.3)	77.1	92.1	93.2	N/A	98.4	16.3
Manual Speaker Segmentation							
C1 (B,I)	95.1 (4.6)	100	88.5	98.3	13.9	100	97.9
C1 (B,D)	66.7 (12.5)	96.9	5.2	66.9	11.8	21.9	12.5
C1 (M,I)	97.0 (4.2)	100	86.5	98.7	61.5	100	97.9
C1 (M,D)	67.5 (9.6)	99.0	6.2	72.0	3.3	6.2	10.4
C2 (B,I)	96.2 (2.6)	96.3	100	96.6	N/A	100	70.4
C2 (B,D)	96.1 (5.8)	96.3	96.3	97.7	N/A	100	33.3
C2 (M,I)	95.8 (7.7)	96.3	96.3	95.7	N/A	100	81.5
C2 (M,D)	98.1 (2.1)	100	100	98.6	N/A	100	48.1

TABLE IV
ROLE RECOGNITION PERFORMANCE FOR C3. THE TABLE REPORTS BOTH THE OVERALL ACCURACY AND THE ACCURACY FOR EACH ROLE. "B" STANDS FOR *BERNOULLI*, "M" STANDS FOR *MULTINOMIAL*, "I" STANDS FOR ROLES *INDEPENDENCE*, AND "D" STANDS FOR ROLES *DEPENDENCE*. THE OVERALL ACCURACY IS ACCOMPANIED BY THE STANDARD DEVIATION σ OF THE PERFORMANCES ACHIEVED OVER THE SINGLE RECORDINGS. THE UPPER PART OF THE TABLE REPORTS THE RESULTS OBTAINED OVER THE OUTPUT OF THE SPEAKER SEGMENTATION, THE LOWER PART REPORTS THE RESULTS OBTAINED OVER THE MANUAL SPEAKER SEGMENTATION.

	all (σ)	PM	ME	UI	ID
Automatic Speaker Segmentation					
C3 (B,I)	46.0 (24.7)	79.6	13.1	41.4	20.3
C3 (B,D)	**46.4** (30.0)	68.7	26.0	32.9	25.7
C3 (M,I)	39.3 (24.9)	67.4	18.0	19.3	25.6
C3 (M,D)	43.7 (31.3)	67.4	28.7	22.0	24.3
Manual Speaker Segmentation					
C3 (B,I)	51.2 (24.2)	83.3	15.9	42.0	29.0
C3 (B,D)	56.0 (33.0)	76.1	37.7	40.6	41.3
C3 (M,I)	43.7 (27.3)	67.4	17.4	39.1	21.7
C3 (M,D)	52.6 (27.6)	76.8	29.0	34.1	33.3

between the actual speaker changes and the changes as detected by the diarization process. The sum of all the misalignments, on average, corresponds to roughly 10% of the recording length, and this is the probable explanation of the performance difference when passing from manual to automatic speaker segmentations.

The rest of the errors are due to limits of the role recognition approach that cannot distinguish between different roles when the associated interaction patterns are too similar. This is true, for example, in the case of the low performance of the IP in corpus C1. The interaction pattern of the IP role is similar to that of the Guest, but the latter has higher *a-priori* probability, so it is usually favored as the output of the recognizer.

A qualitative comparison with other approaches is possible only for some works which use parts of the same data as ours. Both [20] and [21] perform experiments over a subset of the AMI meeting corpus (around 5 h of material). The performance in [20] is around 80%, almost twice as much as our approach over the same data (see Section V). However, as the goal is to detect the two most dominant persons, the probability of assigning each person the correct role is 50%, while it is only 25% in our case. The work in [21] reports a 65% recognition rate of the Project Manager, while our work achieves, over the same role, an accuracy of 79%. Considering that our experiments are performed over the whole AMI meeting corpus, while the experiments of [20] and [21] take into account only a subset of 5 h, our approach seems to be more effective in both cases, though the task is not the same. The work in [18] uses the whole AMI corpus, but it applies a different experimental setup. However, it performs exactly the same task as this work, and the role recognition rate is around 60%.

VI. CONCLUSIONS AND FUTURE WORK

This paper has presented an approach for the automatic recognition of roles in multiparty recordings. The problem of role recognition has been addressed only recently in the literature, but it attracts an increasingly growing interest because it is a key point in the automatic analysis of social interactions [1], [2]. The proposed approach has been tested over roughly 90 h of material, one of the biggest datasets ever used in the literature for this task. To the best of our knowledge, this is the first work that compares the performance of an approach over both *informal* and *formal* roles (see Section II for the difference between the two types of role), showing how the role typology influences the effectiveness of the recognition.

The results show that the recognition accuracy is higher than 85% in the case of broadcast data, and it is around 45% in the case of meeting recordings. There are several possible reasons for such a difference. The first, and probably most important, is that broadcast data include formal roles, while meetings include informal ones. Formal roles are easier to model because they impose constraints on the behavior of people that can be detected, represented, and modeled with probabilistic approaches (like in the case of this work). In contrast, informal roles do not necessarily constrain behavior, and so automatic recognition is more difficult through approaches like the one presented in this work, at least for the aspect of behavior used as role evidence in this work, i.e., *who talks with whom and when*.

The second is that the broadcast data are real, while the meeting data are acted. The meetings do not involve people playing the role they actually have in their life, but volunteers that simulate an artificially assigned role they have never played before. This is likely to reduce significantly the performance of any role recognition method.

In the case of the broadcast data, the performance is sufficient to browse effectively the data (users can quickly find segments corresponding to a given role and the mismatch between the ground truth and the automatic output rarely exceeds a few seconds). In the case of meeting recordings, the approach is effective only to identify the Project Manager. This allows one to effectively follow the progress of the meeting because the PM plays the chairman role as well and, as such, is responsible for following the agenda through her/his interventions.

The main limitation of the current approach is that it does not take into account any sequential information. The role of the person speaking at turn n is likely to have a statistical influence on the role of the person speaking at turn $n+1$. This kind of information could be modeled using probabilistic sequence models (e.g., hidden Markov models), as well as statistical language models (e.g., N-grams). Furthermore, the approach proposed in this work uses only the co-occurence turn-taking patterns as role evidence, while other behavioral cues can be extracted from both audio (e.g., prosodic features) and video (e.g., gestures). Both above limitations will be the subject of future work.

ACKNOWLEDGMENT

The authors would like to thank J. Dines and F. Fleuret for technical support, as well as H. Hung and A. Dielmann for commenting on the draft.

REFERENCES

[1] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image Vis. Comput.*, vol. 27, no. 12, 2009.

[2] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social signal processing: State-of-the-art and future perspectives of an emerging domain," in *Proc. ACM Int. Conf. Multimedia*, 2008, pp. 1061–1070.

[3] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image Vis. Comput.*, to be published.

[4] H. Tischler, *Introduction to Sociology*. Fort Worth, TX: Harcourt Brace College, 1990.

[5] S. Wasserman and K. Faust, *Social Network Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1994.

[6] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.

[7] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, and M. Kronenthal, *et al., The AMI Meeting Corpus: A Pre-Announcement*, ser. Lecture Notes in Computer Science. Berlin, Germany: Springer, 2005, vol. 3869, pp. 28–39.

[8] Y. Liu, "Initial study on automatic identification of speaker role in broadcast news speech," in *Proc. Human Language Technology Conf. NAACL, Companion Volume: Short Papers*, Jun. 2006, pp. 81–84.

[9] S. Maskey and J. Hirschberg, "Automatic summarization of broadcast news using structural features," in *Proc. Eur. Conf. Speech Communication and Technology*, 2003, pp. 1173–1176.

[10] A. Vinciarelli, "Sociometry based multiparty audio recordings summarization," in *Proc .Int. Conf. Pattern Recognition*, 2006, pp. 1154–1157.

[11] A. Vinciarelli and S. Favre, "Broadcast news story segmentation using social network analysis and hidden Markov models," in *Proc. ACM Int. Conf. Multimedia*, 2007, pp. 261–264.

[12] C. Weng, W. Chu, and J. Wu, "Movie analysis based on roles social network," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2007, pp. 1403–1406.

[13] A. Vinciarelli, "Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling," *IEEE Trans. Multimedia*, vol. 9, pp. 1215–1226, 2007.

[14] J. Levine and R. Moreland, "Small groups," in *The Handbook of Social Psychology*, D. Gilbert and G. Lindzey, Eds. Oxford, U.K.: Oxford Univ. Press, 1998, vol. 2, pp. 415–469.

[15] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker, "The rules behind the roles: Identifying speaker roles in radio broadcasts," in *Proc. 17th Nat. Conf. Artificial Intelligence*, 2000, pp. 679–684.

[16] C. Weng, W. Chu, and J. Wu, "Rolenet: Movie analysis from the perspective of social networks," *IEEE Trans. Multimedia*, vol. 11, pp. 256–271, 2009.

[17] F. Pianesi, M. Zancanaro, E. Not, C. Leonardi, V. Falcon, and B. Lepri, "A multimodal annotated corpus of consensus decision making meetings," *J. Lang. Res. Eval.*, vol. 41, no. 3-4, pp. 409–429, 2008.

[18] K. Laskowski, M. Ostendorf, and T. Schultz, "Modeling vocal interaction for text-independent participant characterization in multi-party conversation," in *Proc. 9th ISCA/ACL SIGdial Workshop Discourse and Dialogue*, Jun. 2008, pp. 148–155.

[19] N. Garg, S. Favre, H. Salamin, D. Hakkani-Tür, and A. Vinciarelli, "Role recognition for meeting participants: An approach based on lexical information and social network analysis," in *Proc. ACM Int. Conf. Multimedia*, 2008, pp. 693–696.

[20] D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Predicting the dominant clique in meetings through fusion of nonverbal cues," in *Proc. ACM Int. Conf. Multimedia*, 2008, pp. 809–812.

[21] D. Jayagopi, S. Ba, J. Odobez, and D. Gatica-Perez, "Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues," in *Proc. Int. Conf. Multimodal Interfaces*, 2008, pp. 45–52.

[22] J. Ajmera, "Robust audio segmentation," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, 2004.

[23] J. Dines, J. Vepa, and T. Hain, "The segmentation of multi-channel meeting recordings for automatic speech recognition," in *Proc. Interspeech*, 2006, pp. 1213–1216.

[24] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, pp. 671–680, 1983.

[25] F. Massey Jr., "The Kolmogorov-Smirnov test for goodness of fit," *J. Amer. Statist. Assoc.*, pp. 68–78, 1951.