

Automatic Role Recognition in Multiparty Conversations: An Approach Based on Turn Organization, Prosody, and Conditional Random Fields

Hugues Salamin and Alessandro Vinciarelli

Abstract—Roles are a key aspect of social interactions, as they contribute to the overall predictability of social behavior (a necessary requirement to deal effectively with the people around us), and they result in stable, possibly machine-detectable behavioral patterns (a key condition for the application of machine intelligence technologies). This paper proposes an approach for the automatic recognition of roles in conversational broadcast data, in particular, news and talk shows. The approach makes use of behavioral evidence extracted from speaker turns and applies conditional random fields to infer the roles played by different individuals. The experiments are performed over a large amount of broadcast material (around 50 h), and the results show an accuracy higher than 85%.

Index Terms—Conditional random fields (CRFs), prosody, role recognition, turn organization.

I. INTRODUCTION

ROLES are one of the most important and pervasive aspects of social interaction in at least two main respects [1]. The first is that roles are associated with *shared* expectations that people hold about their own behavior as well as about the behavior of others [2]. Thus, roles contribute to the overall predictability of social interaction, a key condition for making reasonable guesses about others and participating effectively in social exchanges. The second is that roles typically result into “*characteristic behavior patterns*” [2] that can be identified and recognized as such by interaction participants (see the seminal work in [3] and [4]). Thus, roles tend to induce physical, possibly machine-detectable evidences accessible not only to sociological inquiry [5], but also to technological investigation [6], [7].

Manuscript received December 17, 2010; revised April 11, 2011; accepted October 12, 2011. Date of publication October 27, 2011; date of current version March 21, 2012. This work was supported in part by the European Community’s Seventh Framework Programme (FP7/2007-2013) under Grant 231287 (SSPNet), the Swiss National Science Foundation under the National Centre for Competence in Research IM2 (Interactive Multimodal Information Management), and the Scottish Research Council via the Scottish Information and Computer Science Alliance (SICSA). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Qibin Sun.

H. Salamin is with the Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, U.K. (e-mail: hsalamin@dcs.gla.ac.uk).

A. Vinciarelli is with the Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, Scotland, U.K., and also with the Idiap Research Institute CP592, 1920 Martigny, Switzerland (e-mail: vincia@dcs.gla.ac.uk).

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.2011.2173927

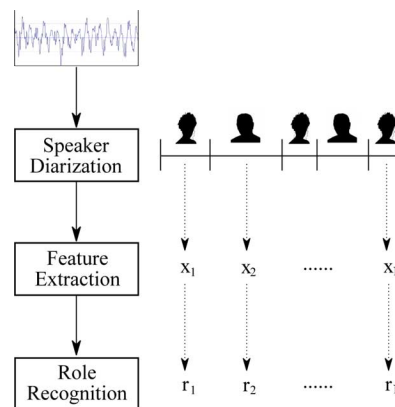


Fig. 1. Approach. The figure shows an overall scheme of the role recognition process.

In light of the above, this work proposes an automatic approach for the recognition of roles in audio recordings of multiparty conversations. The approach includes three main stages (see Fig. 1): the first is the extraction of the *turns*, which are time intervals during which only one person talks, from the raw audio data. This task is performed with an unsupervised speaker diarization approach that does not need to know in advance number and identity of conversation participants [8]. The second stage is the feature extraction, and it represents turns with vectors where each component accounts for a particular measurement. The third stage maps the resulting sequence of vectors into a sequence of roles by using a conditional random field (CRF), which is one of the models most commonly applied for labeling sequences of observations [9]. This corresponds to assigning the speaker of each turn a role, which is the end goal of the entire recognition process.

The reason for using turns as a basic analysis unit is twofold. On the one hand, conversation analysis has shown that the organization of turns (e.g., number of turns, sequence of speakers, distribution of time, and number of turns across speakers) is an important source of socially relevant information [10], typically expressed through ordered sequential patterns such as *preference structures* [11]. On the other hand, the use of turns allows one to extract prosodic information, which is one of the most important channels through which we develop social perceptions about others [12].

Following the taxonomy proposed in [2], the expectations associated to roles (see above) can be expressed as *norms* (explicit

prescriptions about behaviors to be associated to a role), *beliefs* (subjective assessments of what behaviors should be associated to a role), or *preferences* (personal attitudes influencing the behaviors associated to a role). The experiments of this work focus on the first type of expectations and are performed over two corpora of conversational broadcast material (news and talk shows) for a total of roughly 50 h of audio data. The results show that, on average, the proposed approach achieves an accuracy higher than 85%.

This work proposes two main novelties with respect to the *state of the art*. The first is the use of prosody as a cue for role recognition. To the best of our knowledge, prosody has been extensively used for domains like emotion recognition [13], but it has never been used as a role-related cue, even if it is common experience to hear people that change speaking style depending on the particular role they play (e.g., people speak differently when they give a lecture or when they discuss with their friends). Furthermore, the prosody features used in this work are extracted from the current turn only. The features only use information up to the end of the current turn and, to the best of our knowledge, are the first features adequate for online use. The second is, to the best of our knowledge, this is the first work where the experiments are performed over mixed corpora, i.e., over data that share the same role set, but come from multiple sources. This is important to assess how much the approach is robust with respect to variations in the behavioral patterns associated to a given role.

The main technological motivation for the investigation of role recognition is the contribution to the efforts that are being done towards automatic understanding of social interactions (see [14] for an extensive survey). In this respect, role recognition can help to achieve the long term goal of bringing social intelligence into machines. From an application point of view, roles can be used to enrich the content description of multimedia data in retrieval applications, can enhance browsers for data like meeting recordings (see, e.g., the work in [15]), or can allow summarization approaches to identify media segments particularly rich in information [16].

The remainder of this paper is organized as follows. Section II proposes a brief survey of the state of the art, Section III describes the approach in full detail, Section IV reports on experiments and results, and Section V draws some conclusions.

II. STATE OF THE ART

Automatic analysis of social interactions has attracted significant attention in the last few years (see [14] for an extensive survey). In this context, role recognition is one of the problems most commonly addressed and the resulting *state of the art*, while being at a relatively early stage, includes an increasingly wider spectrum of scenarios and approaches. In terms of detectable behavioral patterns, most of the works presented in the literature make use of features related to turn organization (see below). These can be accompanied by other sources of evidence such as lexical choices (e.g., the word distribution in what people say) or movement (e.g., fidgeting).

Role theory relies upon two major concepts: the first are the *behavioral patterns* perceived and recognized as roles by social interaction participants, the second are the *expectations* that shape the behavioral patterns corresponding to roles [2].

Expectations are considered as “role generators” and can be grouped into three broad categories: The *norms* correspond to explicit prescriptions about the behavioral patterns associated to a given role (e.g., during a lecture the teacher is expected to speak while the students are expected to listen). The *beliefs* correspond to subjective choices on how a role should be performed (e.g., people believing that being authoritarian with children is counterproductive expect parents to assume friendly and understanding attitudes). The *preferences* correspond to spontaneous choices based on personality traits or attitudes (e.g., extravert people are expected to look more often than others for social contacts).

In light of the above, the role recognition approaches proposed in the literature can be split into two broad groups. The first includes works aimed at the recognition of roles related to *norms*, the second approaches aimed at the recognition of roles related to *preferences* and *beliefs*. The next two sections provide details about the techniques applied in the two cases.

A. Roles Driven by Norms

The upper part of Table I reports the main aspects of the works dedicated to the recognition of roles for which the expectations are expressed as norms.

The behavioral evidence in [17] corresponds to lexical choices (distribution of uttered terms) and total speaking length. The resulting feature vectors are then mapped into one of the three predefined roles (*journalist*, *guest*, and *anchorman*) using (both individually and in combination) BoosTexter and Maximum Entropy Classifiers. The work in [18] addresses a similar problem (three roles in broadcast news): a hidden Markov model is used to align the sequence of the turns with a sequence of roles, and each turn is represented with the distribution of bigrams and trigrams in the transcription of what is said. The sequence of the roles is modeled with statistical language models. In both [17] and [18], the words at the beginning of each turn appear to be more discriminant than the others. The reason is probably that the beginning of the turn contains a self-introduction of the speaker that often mentions explicitly her role. The work in [22] adopts features that account for turn organization and prosody and maps each person, as detected with a speaker diarization approach, into one of three roles accounting for the general aspects of broadcast data, namely *anchorman*, *journalist*, and *others*. The classification is performed using Gaussian mixture models, k nearest neighbors, or support vector machines (SVM).

The works in [19]–[21] extract automatically social networks from the data in order to assign each person involved in a broadcast recording a different role. The approach in [19] segments the data (audio recordings of news) into turns and then uses the adjacency in the speaker sequence to build a social network. Social-network-based features (e.g., the centrality) are then used to represent each person and map her into one of six predefined roles. In a similar way, the approach in [21] uses the co-presence of two faces (automatically detected and extracted from Hollywood movies) in the same scene as an evidence of direct interaction to build a social network. Features like those applied in [19] are then used to detect the main characters of the movie (the “*leading*” roles) as well as the members of the communities possibly associated with each of the main characters. Finally, the

TABLE I
MAIN RESULTS ON ROLE RECOGNITION PRESENTED IN THE LITERATURE. THE TIME IS EXPRESSED IN HOURS (H) AND MINUTES (M), THE EXPECTATIONS IN TERMS OF *NORMS* (N), *BELIEFS* (B), AND *PREFERENCES* (P)

Ref.	Data	Time	Exp.	Evidence	Approach	Performance
[17]	NIST TREC SDR Corpus (35 recordings, 3 roles)	17h	N	Term distribution, speaking time	BoosTexter, Maximum Entropy Classifier	80.0% of the news stories correctly labeled in terms of role
[18]	TDT4 Mandarin broadcast news (336 shows, 3 roles)	170h	N	Distribution of bigrams and trigrams	Hidden Markov Models	77.0% of the news stories correctly labeled in terms of role
[19]	Radio news bulletins (96 recordings, 6 roles)	25h	N	Turn organization, social networks (centrality, nodes degree, etc.)	Bayesian Classifiers	85% of the data time correctly labeled in terms of role
[20]	Radio news (96 recordings, 6 roles), Talk shows (27 recordings, 6 roles), meetings (138 recordings, 4 roles)	90h	NBP	Turn organization, social networks (centrality, nodes degree, etc.)	Bayesian Classifiers	Up to 85% of the data correctly labeled in terms of role (45% for the meetings)
[21]	Movies and TV shows (13 recordings, 2 roles)	21h	N	Co-occurrence of faces, social networks	Bayesian classifiers	85% to 95% of recognition rate depending on the role
[22]	EPAC Corpus (Broadcast data, 3 roles)	100h	N	Turn organization, prosody	Gaussian Mixture Models, Linear Support Vector Machines, k Nearest Neighbors	92% of role recognition rate
[23]	Meetings (2 recordings, 5 roles)	45m	BP	turn organization	Decision tree	53.0% of segments (up to 60 seconds long) correctly classified
[24]	Mission Survival Corpus (11 recordings, 5 roles)	4h 30m	BP	speaking activity, fidgeting	Support Vector Machines	Up to 70% of analysis windows (10 seconds) correctly classified
[25]	Mission Survival Corpus (11 recordings, 5 roles)	4h 30m	BP	speaking activity, fidgeting	Support Vector Machine	90% of analysis windows (around 10 seconds long) correctly classified
[26]	Mission Survival Corpus (11 recordings, 5 roles)	4h 30m	BP	speaking activity, fidgeting	Influence Model, Support Vector Machines	75% of roles correctly assigned
[27]	AMI Meeting Corpus (138 recordings, 4 roles)	45h	BP	speaking activity, talkspurts	Conditional Random Fields	53% of the time correctly labeled
[28]	AMI Meeting Corpus (138 recordings, 4 roles)	45h	BP	speaking activity, term distribution	Bayesian Classifiers, BoosTexter	75% of the time correctly labeled

approach proposed in [20] uses the turn-taking to build a social affiliation network based on the proximity in time of different speakers. The structure of the network edges is then represented with patterns that are fed to Bayesian classifiers and mapped into roles.

B. Roles Driven by Beliefs and Preferences

The work in [23] applies decision trees to assign meeting participants roles corresponding to different ways of participating in a discussion (i.e., *presenter*, *discussion participator*, *information provider*, *information consumer*, or *Undefined*). The behavioral evidences are extracted from short temporal windows and include the number of speaker changes, the number of participants that have spoken, the number of overlapping speech segments, and the average duration of overlapping speech.

The recognition of *task* roles (*neutral*, *orienteer*, *giver*, *seeker*, and *recorder*) and *socio-emotional* roles (*neutral*, *gate-keeper*, *supporter*, *protagonist*, and *attacker*) as defined in [3] and [4] is the goal of [24]–[26]. In these works, the behavioral evidence is given by speaking activity (e.g., silence versus speaking) and movement (e.g., total amount of fidgeting). The role recognition is performed over short time intervals (2–40 s) that are aligned with a sequence of roles using probabilistic sequential models (e.g., factorial hidden Markov models) or SVMs.

Finally, two works aim at the recognition of roles corresponding to a position in a company in the AMI Meeting Corpus (*Project Manager*, *Marketing Expert*, *Industrial Designer*, and *User Interface Expert*) [27], [28]. The approach in

[27] uses CRFs to align behavioral evidences extracted from short time intervals (e.g., number of times a person talks or a total number of speaking attempts of all meeting participants) with a sequence of roles. The approach in [28] combines lexical choices (distributions of uttered words) and social network features like those applied in [20]. The former features are recognized using the BoosTexter, while the latter uses Bayesian classifiers based on discrete distributions.

III. APPROACH

The approach proposed in this work is depicted in Fig. 1, and it includes three main stages. The first is the extraction of the turns from the raw audio data, the second is the extraction of the features from the turns, and the third is the mapping of the sequence of vectors resulting from the second stage into a sequence of roles. The next three sections describe the stages in more detail.

A. Speaker Diarization

The extraction of the turns is performed with a speaker diarization approach that does not require knowing in advance the number and identity of the speakers. The diarization process is fully described in [8] and will not be further presented here, as it is not the original part of this work. The output of the diarization is a list of triples

$$S = \{(s_1, t_1, \Delta t_1), \dots, (s_N, t_N, \Delta t_N)\} \quad (1)$$

where N is the number of turns extracted by the diarization approach, $s_i \in A = \{a_1, \dots, a_G\}$ is a speaker label, G is the

total number of speakers detected during the diarization, t_i is the starting time of turn i , and Δt_i is its length. The label s_i is not the name of the speaker, but an arbitrary label automatically assigned by the diarization approach. The set A is not defined *a priori*, but it is a result of the diarization process. In general, $G \ll N$ and several turns share the same speaker label. This means that the speaker is the same for the different turns.

If S^* is the ground truth list of triples, i.e., the list where s_i^* corresponds to the real identity of the speaker, t_i^* is the actual time at which turn i starts, Δt_i^* is the actual duration of turn i , G^* is the real number of speakers, and N^* is the real number of turns, then the coherence between S and S^* can be measured in terms of *purity* π as

$$\pi = \sqrt{\pi_c \pi_s} \quad (2)$$

with

$$\pi_c = \sum_{k=1}^G \sum_{l=1}^{G^*} \frac{\Delta t(k)}{T} \left(\frac{\Delta t(lk)}{\Delta t(k)} \right)^2 \quad (3)$$

$$\pi_s = \sum_{l=1}^{G^*} \sum_{k=1}^G \frac{\Delta t^*(l)}{T} \left(\frac{\Delta t(lk)}{\Delta t^*(l)} \right)^2 \quad (4)$$

where $\Delta t(k)$ is the total duration of the S triples for which $s = a_k$, $\Delta t(lk)$ is the total duration of the overlap between S triples for which $s = a_k$, and S^* triples for which $s^* = a_l^*$. The purity is bound between 0 and 1 (the higher the better), and it is one of the most common performance measures in diarization.

B. Feature Extraction

The turn sequence S provides information about *who talks when and how much*. This makes it possible to extract features accounting for the overall organization of turns as well as for the prosodic behavior of each speaker. The turn organization is important because it conveys information about the social actions carried out by different interaction participants [29], typically through “*systematically ordered features*” [10] or appropriate sequences called *preference structures* [11]. The prosody is important because it influences the perception of a large number of socially relevant aspects including competence and expressivity [30], personality [12], and emotional state [13].

Since the earliest works on role theory, both turn organization and prosody have been recognized as one of the main evidences of the roles people play. However, while the turn organization has been extensively used in role recognition literature (see Section II), the prosody has been, to the best of our knowledge, largely neglected. This work tries to fill this gap and proposes using two sets of features for each turn: the first relates to turn organization while the second relates to prosody.

The first set includes features that account for the way an individual interaction participant contributes to the turn organization (i.e., total number of turns for current speaker, time from the beginning of recording to first turn of current speaker, and average time between two turns of current speaker) as well as features that account for how a particular turn contributes to the overall turn organization (i.e., turn duration, time after last turn of the current speaker, and number of unique speakers in the T neighboring turns with $T = 3, 5, \text{ and } 7$). The former features

have the same value for all of the turns where the speaker is the same.

The second set includes the prosodic features, namely the pitch, the first two formants, and the energy and the length of each voiced and unvoiced segment. These measurements are made with Praat [31] over short analysis windows (30 ms) at regular time steps (10 ms) and account for short-term speech aspects. Longer term aspects can be obtained by estimating statistical properties of each feature over the entire turn. In this work, for each feature f , we use the relative entropy

$$H(f) = \frac{-\sum_{f \in F} p(f) \log p(f)}{\log |F|} \quad (5)$$

where F is the set of the f values observed during the turn. The value of $H(f)$ accounts for the overall variability of a feature (the higher the relative entropy, the higher the variability), which is a characteristic that captures the speaking style of a person and influences the social perception that others develop about her [32], [33].

C. Role Recognition

After the feature extraction step, the sequence S of turns is converted into a sequence $X = \{\vec{x}_1, \dots, \vec{x}_N\}$ of observations, where the components of vectors \vec{x}_i correspond to the features described in the previous section. From a statistical point of view, the problem of role recognition can be thought of as finding the sequence of roles $R^* = \{r_1^*, \dots, r_N^*\}$ that satisfies the following equation:

$$R^* = \arg \max_{R \in \mathcal{R}^N} p(R|X) \quad (6)$$

where \mathcal{R} is a predefined set of roles, N is the number of turns, and R is a sequence of N roles.

In this work, the probability $p(R|X)$ is estimated using linear-chain CRFs [9], which is one of the models most commonly applied for labeling observation sequences. The core assumption of this model is that r_{t-i} is conditionally independent of r_{t+j} given r_t and X , for any t and for any i and j greater than 0.

The main advantage of CRFs with respect to other probabilistic sequential models is that they do not require any conditional independence assumption about the observations of X . This is particularly important in this work because some of the features account for long-term dependencies (e.g., the distance with respect to the last turn of the current speaker) and others have the same value for all of the turns of a certain speaker (e.g., the number of turns for the current speaker). In both cases, models based on the assumption that the observations are conditionally independent given an underlying variable (e.g., hidden Markov models) would not be appropriate.

CRFs are undirected graphical models, thus the probability distribution they correspond to can be expressed as a product of functions called *potentials* [34]. The arguments of each potential are the variables represented by the nodes of a clique in the graph underlying the probability distribution. In linear-chain CRFs, the maximal cliques are pairs of nodes corresponding to adjacent elements in the sequence of the labels. In the case of this work, the pairs have the form $\{r_t, r_{t+1}\}$ and include the roles assigned to two consecutive turns. In addition, this work

considers potentials that are function of the role of an individual turn as well (see below).

The following assumptions have been made about the potentials to make the model tractable.

- 1) The potential over $\{r_t, r_{t+1}\}$ depends only on r_t and r_{t+1} .
- 2) The potential over $\{r_t\}$ depends only on r_t and \vec{x}_t .
- 3) The potentials are the same for all t .
- 4) The potentials are never zero.

This first three assumptions mean that the marginal distribution for r_t is fully determined by r_{t-1} , r_{t+1} , and x_t . The fourth assumption means that every role assignment has a probability strictly greater than zero. This last assumption is important in practice because it allows the product of potentials to be replaced by the exponential of a sum as follows [34]:

$$p(R|X) = \frac{\exp\left(\sum_{t=1}^N f_1(r_t, \vec{x}_t) + \sum_{t=1}^{N-1} f_2(r_t, r_{t+1})\right)}{Z(X)}$$

$$Z(X) = \sum_{R \in \mathcal{R}^N} \exp\left(\sum_{t=1}^N f_1(r_t, \vec{x}_t) + \sum_{t=1}^{N-1} f_2(r_t, r_{t+1})\right)$$

where $Z(X)$ is called *partition function* and is simply a normalization constant, and f_1 and f_2 represent potentials having as argument only one role assignment r_t or a pair of consecutive role assignments $\{r_t, r_{t+1}\}$. The potentials have been represented as a linear combination of simpler terms called *feature functions*. In this work, the feature functions used for f_1 are

$$f_{r,i}(r_t, \vec{x}_t) = \begin{cases} x_t^{(i)}, & \text{if } r_t = r \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

where $x_t^{(i)}$ is the i th component of \vec{x}_t . This family of feature functions can capture linear relations between a role and an observation $x_t^{(i)}$. For f_2 , the feature functions applied in this work are

$$f_{r,r'}(r_t, r_{t+1}) = \begin{cases} 1, & \text{if } r_t = r \text{ and } r_{t+1} = r' \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

In summary, linear-chain CRFs estimate the *a posteriori* probability $p(R|X)$ of a role sequence as follows:

$$p(R|X, \alpha) = \frac{1}{Z(X)} \exp\left(\sum_{t=1}^N \sum_{r \in \mathcal{R}} \sum_i \alpha_{r,i} f_{r,i}(r_t, \vec{x}_t) + \sum_{t=1}^{N-1} \sum_{(r,r') \in \mathcal{R}^2} \alpha_{r,r'} f_j(r_t, r_{t+1})\right). \quad (9)$$

The weights $\alpha_{r,i}$ of the feature functions of form $f_{r,i}(R, X)$ account for how much the value of a given feature is related to a particular role. The weights of the feature functions of form $f_{r,r'}(R, X)$ account for how frequent it is to find role r followed by role r' .

Given a training set $\{(X_j, R_j)\}$ of labelled sequences, the weights α are learned using the maximum-likelihood approach

$$\alpha = \arg \max_{\alpha^*} \sum_j \log p(R_j | X_j, \alpha^*). \quad (10)$$

In the case of CRFs, this maximization can be accomplished using gradient ascent techniques.

TABLE II
ROLE DISTRIBUTION. THE PERCENTAGE OF TIME EACH ROLE ACCOUNTS FOR IN THE TWO CORPORA

Role	AM	SA	GT	HR	WM	IP
C1	41.2%	5.5%	34.8%	7.1%	6.3%	4.0%
C2	17.3%	10.3%	64.9%	4.0%	1.7%	0.0%

IV. EXPERIMENTS AND RESULTS

Here, we describe in detail the data and the recognition results obtained during the experiments of this work.

A. The Data

The experiments of this work have been performed over two corpora of radio broadcast material. The first corpus, referred to as C1, contains 96 news bulletins (19 h of material in total). The second corpus, referred to as C2, contains 27 talk shows (27 h of material in total). C1 includes all news bulletins delivered by Radio Suisse Romande (the Swiss national French-speaking broadcast service) in February 2005, while C2 includes all editions of a popular radio talk show (*Forum*) broadcast during the same period. Compared with television material, radio data provide much less information, both from behavioral (e.g., no facial expressions, gestures, and body movements) and multimedia (e.g., no shot segmentation or captions) points of view. On the other hand, if a method works on radio data, it can be extended to video data as well (using only the audio stream), possibly in combination with approaches dealing with visual information, while the reverse is not always possible.

The role set \mathcal{R} includes, for both corpora, the following roles: *Anchorman* (AM), *Guest* (GT), *Second Anchorman* (SA), *Interview Participant* (IP), *Headline Reader* (HR), and *Weather Man* (WM). The same role name does not correspond necessarily to the same expectations in the two corpora. For example, the AM is expected to inform in news and to entertain in talk shows. This is likely to lead to different behavior patterns for the same role in the two corpora. Table II reports the percentage of time each role accounts for in both C1 and C2. The purity of the speaker diarization (see Section III) is, on average, 0.79 for C1 and 0.81 for C2.

The number of roles is smaller than the number of people talking in each recording: the average number of speakers is 12 in C1 and 30 in C2. Certain roles are played only by one person per recording (e.g., there is only one AM per talk show), while others can be played by several persons (e.g., there are several GT per news bulletin). Furthermore, the same person can play different roles in different recordings. Hence, there is no one-to-one mapping between speaker identity and role (this is the reason why there is no attempt to recognize the speakers).

B. Recognition Results

The recognition experiments have been performed using a k -fold approach ($k = 5$): each corpus has been split into k disjoint subsets and, iteratively, each one of these has been used as a test set while the others have been used as training set. The k -fold approach allows one to use the entire dataset at disposition for testing purposes while still keeping a rigorous separation between training and test data [35].

TABLE III
ACCURACY VALUES WHEN USING ONLY PROSODIC FEATURES (P), ONLY TURN-ORGANIZATION FEATURES (T), OR THE COMBINATION OF THE TWO (PT). THE UPPER PART OF THE TABLE REPORTS THE RESULTS ACHIEVED OVER THE TURNS EXTRACTED AUTOMATICALLY (A), WHILE THE LOWER PARTS REPORTS THOSE ACHIEVED OVER THE MANUAL SPEAKER SEGMENTATION (M)

Corpus	P	T	PT
C1 (A)	83.0%	89.7%	89.3%
C2 (A)	69.5%	84.2%	87.0%
C1+C2 (A)	68.1%	86.4%	86.7%
C1 (M)	87.1%	99.1%	99.1%
C2 (M)	76.2%	96.9%	96.2%
C1+C2 (M)	75.8%	96.6%	96.5%

Table III reports the overall recognition results obtained over C1 and C2 separately, as well as on their union. The performance is reported in terms of accuracy, which is the percentage of time correctly labeled in terms of role in the test set. The upper part of the table shows the recognition results when using an automatic speaker diarization, while the lower part reports the results when segmenting the audio data into turns manually. In the former case, the segmentation is affected by errors, while in the latter case it corresponds to the actual turns in data. The performance over the manual segmentation is higher than 95% for all of the corpora and this seems to suggest that the features adopted in this work capture, at least in part, the behavior patterns associated to the roles. The performance loss when moving to the automatic speaker segmentation is typically slightly lower than 10%. The main reason is that speaker changes are detected with a certain delay (1–2 s) and the accumulation of these misalignments sums up, on average, to roughly 8% of the recordings length. In the case of C1, the best accuracy on the automatic segmentation is 91.8%. For C2, the best accuracy on the automatic segmentation is 93.0%.

The performance is reported using only prosodic features (column P), only turn-organization features (column T), and the combination of the two (column $P + T$). Prosodic features appear to perform less than the others, but still have a satisfactory performance. This is important because prosody can be extracted online. Hence, prosody allows one to recognize roles as the interaction progresses without waiting for having at disposition the entire recording (as required in the case of turn-organization features). The above might represent an advantage in scenarios like automatic captioning or computer-mediated communication where automatic analysis is expected to work online.

The performance of prosodic features is lower compared with the others, but the high performance of turn-organization features on the manual segmentation (see accuracies higher than 96%), close to the maximum possible accuracy on the automatic speaker segmentation of C1 (89.7% versus 91.8%), might actually hide the contribution of prosody in three cases out of four. Not surprisingly, the combination of prosody and turn organization leads to statistically significant improvements on C2 alone, where the turn-organization features show the lowest accuracy.

The recognition experiments have been performed not only over C1 and C2 separately, but also over their union. As the results are comparable to those obtained over C1 and C2 individually, the role recognition approach seems to be robust with

TABLE IV
PURITY OF THE ROLE ASSIGNMENT, I.E., THE COHERENCE BETWEEN SPEAKER LABEL AND ROLE

Corpus	P	T	PT
C1 (A)	0.84	0.94	0.94
C2 (A)	0.84	0.93	0.93
C1 (M)	0.93	0.99	0.99
C2 (M)	0.84	0.98	0.98

TABLE V
ROLE ACCURACY FOR C1

	AM	SA	GT	HR	WM	IP
P (A)	66.1%	0.0%	60.5%	88.2%	90.2%	0.0%
T (A)	96.5%	11.7%	94.1%	97.8%	96.0%	13.7%
PT (A)	96.5%	11.6%	94.1%	97.4%	93.5%	12.3%
P (M)	94.7%	77.8%	93.3%	100%	93.9%	31.2%
T (M)	99.9%	96.6%	99.0%	100%	99.1%	93.0%
PT (M)	99.7%	96.6%	98.8%	100%	97.9%	88.7%

respect to a higher variability in the behavioral patterns through which roles are played.

The results of this work can be compared with those obtained in [20], where the experiments have been conducted over the same data (C1 and C2) and the same experimental protocol has been used. The approach proposed here differs from the previous one [20] in several respects. This work uses a probabilistic sequential model taking into account the sequence of the roles in a conversation, while the previous one uses a social network to represent the overall structure of the turns. This work assigns the roles turn by turn, while the previous one assigns the roles person by person. Furthermore, this work uses prosodic features and turn organization, while the previous one is based only on turn organization. The best results reported for C1 and C2 in [20] are 82.4% and 87.8%, respectively. In this work, the best performances are 89.3% for C1 and 87.0% for C2. In the case of C1, taking into account sequential aspects and prosody produces a statistically significant improvement (the error rate is decreased by 39% with a p -value lower than 0.0001), while in the case of C2 the difference is not statistically significant (the error rate increases by 9%, but the p -value is 0.46).

As a role is assigned to each turn, the same person can be assigned multiple roles as the conversation evolves. This is a desirable characteristic of the approach because, in many scenarios, individuals can play different roles in the same conversation. However, in the scenario used in this work, each person plays only one role and the approach should be robust with respect to this aspect. Table IV reports the purity of the role assignment, i.e., the coherence between speaker labels and roles (see Section III for a definition of the purity). For T and PT features, the purities are always higher than 0.9, and this clearly suggests that the same person tends to be assigned always the same role.

Tables V–VII provide the performance for each role separately. Some roles are recognized with high accuracy in all of the corpora (e.g., WM, HR, and GT), while others show significant differences depending on the data. The most likely explanation is that some roles correspond to characteristic patterns in all of the cases, while others do not. Furthermore, there seems to be

TABLE VI
ROLE ACCURACY FOR C2

	AM	SA	GT	HR	WM
P (A)	43.1%	9.7%	92.1%	94.7%	0.9%
T (A)	72.6%	85.8%	92.6%	95.0%	13.3%
PT (A)	76.4%	85.8%	92.9%	95.0%	41.5%
P (M)	70.1%	15.4%	94.6%	96.3%	0.0%
T (M)	99.4%	95.4%	98.8%	96.3%	81.5%
PT (M)	98.2%	85.3%	97.5%	100%	74.1%

TABLE VII
ROLE ACCURACY FOR THE UNION OF C1 AND C2

	AM	SA	GT	HR	WM
P (A)	37.8%	0.6%	72.0%	83.6%	67.8%
T (A)	92.6%	16.2%	93.1%	94.8%	72.6%
PT (A)	91.2%	18.2%	92.3%	92.5%	74.3%
P (M)	71.1%	34.2%	92.7%	98.4%	69.1%
T (M)	99.7%	92.4%	99.3%	100%	84.6%
PT (M)	99.0%	90.8%	99.0%	100%	87.0%

a tendency, on average, to achieve higher recognition accuracy for roles that have higher *a priori* probability (see Table II).

V. CONCLUSION

This paper has addressed the problem of role recognition in conversational broadcast data. The proposed approach uses turns as basic analysis unit and extracts features accounting for the organization of turns as well as for the prosody of speakers. The experiments have been performed over two corpora containing news and talk-shows and the roles are associated to norms, i.e., explicit prescriptions about the actual behavior that must be associated to the role. The performance of the proposed approach can be considered satisfactory as most of the error seems to depend not on the modeling of the roles, but on the errors of the speaker diarization approach, i.e., of the step necessary to extract the turns from the raw audio data.

Most of the role recognition approaches presented in the literature (see Section II) use the organization of turns as an evidence of the roles being played and this work is no exception. However, to the best of our knowledge, this is one of the very first works that combine the organization of turns with prosody. This is important because prosodic behavior (the way a person talks) influences to a significant extent the social perception of people [32], [33] and it is likely to account for the role someone plays.

The results show that the prosodic features are less effective than those related to turn organization. However, the combination of the two types of features leads to statistically significant improvements for the corpus where turn organization features have the lowest performance. This seems to suggest that more conclusive results could be obtained only by using data where turns (or any other evidence at disposal) are not effective enough to hide the positive effect of other features.

The performance has been compared with results previously obtained over the same data with different approaches [20]. The comparison suggests that taking sequential aspects of a conversation into account (in particular how roles tend to follow one another) leads to significant improvements.

This work has addressed the role recognition problem in machine intelligence terms, i.e., by trying to maximize the accuracy

of the approach. No attempt has been made to explain what are the behavioral patterns the roles correspond to. An unsupervised analysis of the feature vectors extracted in this work might show what are the most salient behavioral aspects of each role. Such an approach might be of help when trying to identify characteristic behavior patterns for roles associated to preferences or beliefs, in scenarios that are less constrained and where the roles might be more difficult to define *a priori*.

REFERENCES

- [1] J. Scott and G. Marshall, Eds., *Dictionary of Sociology*. Oxford, U.K.: Oxford Univ., 2005.
- [2] B. Biddle, "Recent developments in role theory," *Ann. Rev. Sociol.*, vol. 12, pp. 67–92, 1986.
- [3] R. Bales, "A set of categories for the analysis of small group interaction," *Amer. Sociolog. Rev.*, vol. 15, no. 2, pp. 257–263, 1950.
- [4] P. Slater, "Role differentiation in small groups," *Amer. Sociolog. Rev.*, vol. 20, no. 3, pp. 300–310, 1955.
- [5] H. Tischer, *Introduction to Sociology*. Orlando, FL: Harcourt Brace, 1990.
- [6] C. Song, Z. Qu, N. Blumm, and A. Barabasi, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.
- [7] A. Vinciarelli, "Capturing order in social interactions," *IEEE Signal Process. Mag.*, vol. 26, no. 5, pp. 133–137, May 2009.
- [8] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2004, pp. 411–416.
- [9] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. Int. Conf. Mach. Learning*, 2001, pp. 282–289.
- [10] H. Sacks, E. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.
- [11] J. Bilmes, "The concept of preference in conversation analysis," *Language Soc.*, vol. 17, no. 2, pp. 161–181, 1988.
- [12] C. Nass and S. Brave, *Wired for Speech*. Cambridge, MA: MIT Press, 2005.
- [13] K. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, vol. 40, no. 1–2, pp. 227–256, 2003.
- [14] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image Vis. Computing J.*, vol. 27, no. 12, pp. 1743–1759, 2009.
- [15] L. Matena, A. Jaimes, and A. Popescu-Belis, "Graphical representation of meetings on mobile devices," in *Proc. 10th Int. Conf. Mobile Human Comput. Interaction*, 2008, pp. 503–506.
- [16] A. Vinciarelli, "Sociometry based multiparty audio recordings summarization," in *Proc. Int. Conf. Pattern Recognit.*, 2006, vol. 2, pp. 1154–1157.
- [17] 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. Artif. Intell.*, 2000, pp. 679–684.
- [18] Y. Liu, "Initial study on automatic identification of speaker role in broadcast news speech," in *Proc. Human Language Technol. Conf. NAACL*, Jun. 2006, pp. 81–84.
- [19] A. Vinciarelli, "Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling," *IEEE Trans. Multimedia*, vol. 9, no. 6, pp. 1215–1226, Jun. 2007.
- [20] H. Salamin, S. Favre, and A. Vinciarelli, "Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1373–1380, Jul. 2009.
- [21] C. Weng, W. Chu, and J. Wu, "Rolenet: Movie analysis from the perspective of social networks," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 256–271, Feb. 2009.
- [22] B. Bigot, I. Ferrané, J. Pinquier, and R. André-Obrecht, "Speaker role recognition to help spontaneous conversational speech detection," in *Proc. Int. Workshop Searching Spontaneous Conversational Speech*, 2010, pp. 5–10.
- [23] S. Banerjee and A. Rudnicky, "Using simple speech based features to detect the state of a meeting and the roles of the meeting participants," in *Proc. Int. Conf. Spoken Language Process.*, 2004.
- [24] M. Zancanaro, B. Lepri, and F. Pianesi, "Automatic detection of group functional roles in face to face interactions," in *Proc. Int. Conf. Multimodal Interfaces*, 2006, pp. 47–54.

- [25] F. Pianesi, M. Zancanaro, E. Not, C. Leonardi, V. Falcon, and B. Lepri, "Multimodal support to group dynamics," *Personal Ubiquitous Computing*, vol. 12, no. 3, pp. 181–195, 2008.
- [26] W. Dong, B. Lepri, A. Cappelletti, A. Pentland, F. Pianesi, and M. Zancanaro, "Using the influence model to recognize functional roles in meetings," in *Proc. 9th Int. Conf. Multimodal Interfaces*, Nov. 2007, pp. 271–278.
- [27] 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 Dialogue*, Jun. 2008, pp. 148–155.
- [28] 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.
- [29] G. Psathas, *Conversation Analysis: The Study of Talk-in-Interaction*. Newbury Park, CA: Sage, 1995.
- [30] P. Ekman, W. Friesen, M. O'Sullivan, and K. Scherer, "Relative importance of face, body, and speech in judgments of personality and affect," *J. Personality Social Psychol.*, vol. 38, no. 2, pp. 270–277, 1980.
- [31] P. Boersma, "Praat, a system for doing phonetics by computer," *GLOT Int.*, vol. 5, no. 9/10, pp. 341–345, 2002.
- [32] D. Addington, "The relationship of selected vocal characteristics to personality perception," *Commun. Monographs*, vol. 35, no. 4, pp. 492–503, 1968.
- [33] G. Ray, "Vocally cued personality prototypes: An implicit personality theory approach," *Commun. Monographs*, vol. 53, no. 3, pp. 266–276, 1986.
- [34] D. Koller and N. Friedman, *Probabilistic Graphical Models*. Cambridge, MA: MIT Press, 2009.
- [35] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. Joint Conf. Artif. Intell.*, 1995, vol. 14, pp. 1137–1145.



Hugues Salamin received the M.Sc. degree in mathematics from the Swiss Federal Institute of Technology, Zurich, Switzerland, in 2007. He is currently working toward the Ph.D. degree at the School of Computing Science, University of Glasgow, Glasgow, U.K.

During his Ph.D. studies, he has been a Visiting Intern with the Institute for Creative Technology, Los Angeles, CA. His current research interests is the development of an automatic system for understanding nonverbal behavior in human interactions.



Alessandro Vinciarelli is a Lecturer with the University of Glasgow, Glasgow, U.K., and a Senior Researcher with the Idiap Research Institute, Martigny, Switzerland. His main research interest is Social Signal Processing, the new domain aimed at bringing social intelligence in computers. He is coordinator of the FP7 Network of Excellence SSPNet and is, or has been, Principal Investigator of several national and international projects. He has authored and coauthored more than 60 publications, including one book and 21 journal papers.

Dr. Vinciarelli has initiated and organized a large number of international workshops, including the International Workshop on Socially Intelligent Surveillance and Monitoring and the International Workshop on Social Signal Processing. He is cocchair of the IEEE Technical Committee on SSP and is an associate editor of the IEEE SIGNAL PROCESSING MAGAZINE for the social sciences. Furthermore, he is founder of a knowledge management company (KleweI) recognized with several national and international prizes.