Automatic Role Recognition in Multiparty Recordings Using Social Networks and Probabilistic Sequential Models

S. Favre[1,2], A. Dielmann[1], and A. Vinciarelli[1,2]
[1]Idiap Research Institute - CP592, 1920 Martigny, Switzerland
[2]Ecole Polytechnique Federale de Lausanne - 1015 Lausanne, Switzerland
{sfavre, adielman, vincia}@idiap.ch

ABSTRACT

The automatic analysis of social interactions is attracting significant interest in the multimedia community. This work addresses one of the most important aspects of the problem, namely the recognition of roles in social exchanges. The proposed approach is based on Social Network Analysis, for the representation of individuals in terms of their interactions with others, and probabilistic sequential models, for the recognition of role sequences underlying the sequence of speakers in conversations. The experiments are performed over different kinds of data (around 90 hours of broadcast data and meetings), and show that the performance depends on how formal the roles are, i.e. on how much they constrain people behavior.

Categories and Subject Descriptors: H.3.1 [Content Analysis and Indexing]. **General Terms:** Experimentation. **Keywords:** Social Network Analysis, Role Recognition, HMMs, Statistical Language Models.

1. INTRODUCTION

The multimedia community is making significant efforts towards the automatic analysis of social interactions in audio and video recordings (see [11] for an extensive survey). This work considers one of the key aspects of the problem, i.e. the recognition of *roles* in multiparty recordings. Roles are not only a universal aspect of social exchanges, because people play roles each time they interact with others [10], but they can also be useful in several applications, e.g. in media browsers, in summarization and in Information Retrieval. Indeed, in media browsers, the role of the person speaking at a given moment can help users to quickly identify segments of interest. In summarization, the role can be used as a criterion to select representative segments of the data. In Information Retrieval, the role can be used as an index to enrich the content description of the data.

The core idea of the approach we propose is that *the sequence of speakers talking during a conversation is the observable, machine detectable, evidence of an underlying, hidden, sequence of roles.* Correspondingly, the approach includes three main steps: the first is the *Speaker Diarization* and splits the multiparty recordings into *turns*, i.e. single speaker intervals (see Section 2.1). The second is the *Feature Extraction* and applies Social Affiliation Networks [12] to convert the sequence of turns into a sequence of feature vectors accounting for how each speaker interacts with others (see Section 2.2). The third is the actual *Role Recognition*, where the sequence of feature vectors is aligned with a sequence of roles using Hidden Markov Models (HMM) [8] and Statistical Language Models (SLM) [9] (see Section 2.3).

The experiments are performed over three different corpora for a total of around 90 hours of material (see Section 3.1). The results show that the approach is particularly suitable for the recognition of *formal* roles, i.e. those that correspond to specific functions in a given interaction setting (e.g. the moderator in a debate) and impose more or less rigorous constraints on the behavior of people [5]. *Informal* roles, i.e. those that correspond to a position in a specific social system (e.g., the manager in a company) and do not impose constraints on the behavior of people [5], are harder to model, but still recognized with a performance singificantly higher than chance. To the best of our knowledge, there is only one work reporting results over a larger dataset [6], but it considers only formal roles, thus the results are less exhaustive than those presented here.

The main novelties and distinctive aspects of this article with respect to the *state-of-the-art* are as follows:

- This is the first work, to the best of our knowledge, that *provides a quantitative measure of how formal a role set is, i.e. of how much the roles under consideration constrain the interaction behavior of people.* This is important to assess how effectively a role recognition approach can work in different interaction settings.

- This is the first work, to the best of our knowledge, that *assesses how diverse are role recognition approaches based on probabilistic sequential models and on Bayesian classifiers.* This is important in view of the combination of different role recognition techniques.

- To the best of our knowledge, *the dataset used in this work is the only one that includes different interaction settings and different role sets.* This is important to assess how easily an approach can be ported from one interaction setting to another.

MM'09 October 19-24, 2009, Beijing, China.

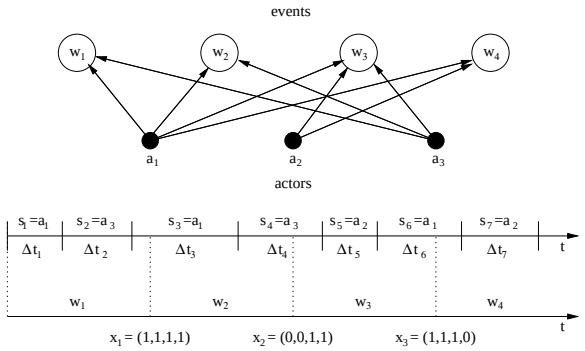

Figure 1: Social Affiliation Network extraction. The events correspond to the segments w_j and the actors are linked to the events when they talk during the corresponding segments. The actors are represented using vectors $\vec{x}_i$ where the components account for the links between actors and events.

The rest of this paper is organized as follows: Section 2 describes the role recognition approach, Section 3 presents experimental results, and Section 4 draws some conclusions.

2. THE ROLE RECOGNITION APPROACH

The next three sections describe the three main steps of the recognition process.

2.1 Speaker Diarization

The goal of a speaker diarization process is to segment an audio recording into *turns*, i.e. time intervals during which there is only one person speaking. The experiments of this work are performed on one hand over broadcast data and, on the other hand, on meeting recordings. Correspondingly, two diarization approaches have been applied to cope with the characteristics of each type of data. The diarization techniques applied to broadcast data and meeting recordings are fully described in [1] and [3], respectively.

2.2 Feature Extraction

The speaker diarization process converts each recording into a sequence of turns $S = \{(s_i, \Delta t_i)\}$, where $i = 1, \ldots, N$, s_i is the speaker label corresponding to the voice detected in the i^{th} turn, and Δt_i is the duration of the i^{th} turn. The label s_i belongs to the set $A = \{a_1, \ldots, a_G\}$ of G unique speaker labels as provided by the speaker diarization process. The turn sequence S is used to build a Social Affiliation Network (SAN), a graph with two types of nodes (*actors* and *events*) where links are not allowed between nodes of the same type (see Figure 1) [12]. In our experiments, the actors correspond to the speakers as detected in the diarization process, and the events correspond to uniform non-overlapping segments spanning the whole length of the recording (see lower part of Figure 1). The rationale is that actors participating in the same events (i.e. participants speaking during the same time intervals) are likely to interact with each other. Therefore, the SAN extracts the evidence of interactions in terms of: *who talks to whom and when.*

One of the main advantages of this representation is that each actor a_i can be represented with a D-dimensional vector $\vec{x}_i$, where the component j accounts for the participation of a_i in event j. The j^{th} component is set to 1 if the speaker talks during the j^{th} segment, and to 0 otherwise (see bottom of Figure 1). A further component is added corresponding to the fraction of time τ_i speaker i talks for during a given recording. The resulting vector $\vec{z}_i = (\vec{x}_i, \tau_i)$ has dimension $D+1$. The dimensionality of the feature vectors is reduced through Principal Component Analysis (PCA) [2]. The amount of variance to be retained after PCA is arbitrarily set to 70%, while D is set through crossvalidation during the experiments. In some cases, D influences significantly the role recognition performance as it defines the events during which interactions are captured. This effect can be measured by comparing the performances obtained when D is selected through crossvalidation and when D is selected maximizing the performance over the test set. The latter procedure overestimates the performance, but it gives an idea of how much D actually influences the role recognition accuracy (see Section 3).

2.3 Role Recognition

The application of PCA to the $\vec{z}_i$ feature vectors results into M-dimensional projections $\vec{y}_i$, where $M < D+1$. Therefore, each recording can be represented through a sequence $Y = (\vec{y}_1, \ldots, \vec{y}_N)$, where N is the number of turns detected at the speaker diarization step, and $\vec{y}_i$ is the vector representing the speaker talking at turn i.

The role recognition can be thought of as finding the role sequence R^* satisfying the following equation:

$$R^* = \arg \max_{R \in \mathcal{R}^N} p(Y|R)p(R), \quad (1)$$

where $R = (r_1, \ldots, r_N)$ is a sequence of roles of length N, $r_i \in \mathcal{R}$ ($\mathcal{R}$ is a predefined set of roles), and $\mathcal{R}^N$ is the set of all role sequences of length N. In intuitive terms, the above equation says that R^* is the sequence of roles that better explains (in terms of a-posteriori probability) the sequence of turns actually observed during a conversation.

In our experiments, the likelihood $p(Y|R)$ is estimated with a fully connected, ergodic, HMM [8] where each state corresponds to a role $r \in \mathcal{R}$. Each state can be reached from any other state, meaning that transitions between any pair of roles are allowed. The emission probability function associated to each state are Gaussians.

The *a-priori* probability $p(R)$ is estimated using a n-gram ($n \geq 1$) Statistical Language Model [9]:

$$p(R) = \prod_{k=1}^{N} p(r_k | r_{k-1}, r_{k-2}, \ldots, r_{k-n+1}). \quad (2)$$

HMMs and SLMs have been implemented with two publicly available packages, the Hidden Markov Model Toolkit, and the SRI Language Model Toolkit.

3. EXPERIMENTS AND RESULTS

This section outlines the experimental setup and the automatic role recognition results.

3.1 Data and Roles

The experiments of this work are performed over three different corpora that will be referred to as C1, C2 and C3 (Table 1 shows their main characteristics). The roles of C1 are *Anchorman* (AM), *Second Anchorman* (SA), *Guest* (GT), *Interview Participant* (IP), *Headline Person* (HP),

DB	recs.	setting	tot. t	avg. t	avg. G
C1	96	news	18h 56m	11m 50s	12
C2	27	talk-show	27h 00m	1h 00m	30
C3	137	meeting	45h 38m	19m 50s	4

Table 1: Corpora. The table reports the main characteristics of the corpora used in the experiments. From left to right: number of recordings, interaction setting, total time, average recording length, average number of participants. Note that the length is the same (one hour) for all recordings in C2, and the number of participants is constant (four) in C3. In all other cases, the figures change from one recording to the other.

and *Weather Man* (WM). Roles with the same name are played in C2 (with the exception of IP that appears only in C1), but they correspond to different functions because the interaction setting changes singificantly between C1 and C2 (e.g., AM are expected to inform in news and to entertain in talk-shows). The roles of C3 are *Project Manager* (PM), *Marketing Expert* (ME), *User Interface Expert* (UI), and *Industrial Designer* (ID).

The effectiveness of the diarization processes (see Section 2.2) is measured with the *Purity* π:

$$\pi = \left[\sum_{k=1}^{N_c} \sum_{l=1}^{N_s} \frac{n_k}{N} \frac{n_{lk}^2}{n_k^2} \sum_{l=1}^{N_s} \sum_{k=1}^{N_c} \frac{n_l}{N} \frac{n_{lk}^2}{n_l^2} \right]^{\frac{1}{2}}, \qquad (3)$$

where N is the total number of feature vectors extracted from a audio recording, 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, n_k is the number of feature vectors in detected voice k, and n_l is the number of vectors belonging to speaker l. The purity ranges between 0 and 1, the higher the better. The average purity is 0.82 for C1, 0.78 for C2 and 0.99 for C3.

3.2 Role Recognition Results

The experiments have been performed with a leave-one-out approach [2]: Each recording is iteratively used as test set while all others are used as training set. This approach ensures a rigorous separation between training and test set while allowing one to perform tests over the whole dataset at disposition. The experiments involve two hyperparameters, the number D of events in the Social Affiliation Network and the amount of variance retained after applying PCA. D is set through crossvalidation (the value resulting into the best performance over the training set is retained for testing), the variance value has been set *a-priori* to 70% and no crossvalidation has been performed.

Tables 2 and 3 report the recognition performance in terms of *accuracy*, i.e. percentage of time correctly labeled in terms of role. For each corpus, the first row (HMM) shows the results when using only HMMs, the others show the accuracy achieved with language models of increasing order (HMM+n-gram). For each corpus, the last row reports, for comparison purposes, the performance achieved with an approach (Bayes) previously proposed by the authors and based on a Bayesian classifier [4].

	all (σ)	AM	SA	GT	IP	HP	WM
Results over C1							
HMM	74.2 (8.8)	97.8	9.5	65.9	38.0	61.1	63.0
HMM + 1-gram	77.7 (11.6)	93.0	9.6	83.4	7.9	59.1	80.5
HMM + 2-gram	79.7 (12.6)	95.5	12.2	82.6	25.7	63.5	80.4
HMM + 3-gram	79.7 (9.3)	97.8	10.3	81.4	25.7	59.5	78.0
Bayes	82.5 (6.9)	98.0	3.6	91.8	8.0	64.6	79.9
Results over C2							
HMM	71.7 (7.2)	74.4	91.9	69.8	N/A	62.1	74.6
HMM + 1-gram	83.7 (6.7)	74.5	92.0	90.3	N/A	58.4	19.0
HMM + 2-gram	82.4 (7.4)	74.7	91.3	88.0	N/A	51.1	24.5
HMM + 3-gram	86.1 (6.8)	74.4	91.9	92.0	N/A	72.8	30.5
Bayes	82.6 (6.9)	75.0	88.3	91.6	N/A	18.3	6.7

Table 2: Role recognition performance on C1 and C2. The table reports both the overall accuracy and the accuracy for each role. The overall accuracy is accompanied by the standard deviation σ of the performance achieved over the single recordings.

	all (σ)	PM	ID	ME	UI
HMM	44.4 (26.8)	63.4	22.6	28.4	40.1
HMM + 1-gram	40.7 (24.4)	70.6	12.4	16.1	32.0
HMM + 2-gram	48.0 (25.9)	63.3	28.3	35.8	34.6
HMM + 3-gram	46.9 (24.9)	61.8	25.0	39.4	33.8
Bayes	43.5 (23.9)	75.3	15.1	15.1	40.0

Table 3: Role recognition on C3. The table reports both the overall accuracy and the accuracy for each role. The overall accuracy is accompanied by the standard deviation σ of the performance achieved over the single recordings.

Even if the training material at disposition is sufficient to train models of order up to 6, no performance improvements are observed for $n > 3$. This seems to suggest that higher order dependences do not bring any information and the role observed at turn k depends at most on the last two preceding roles.

The performance tends to be higher for those corpora where the *Perplexity* PP of the language models is lower:

$$PP = [\prod_{k=1}^{N} p(r_k | r_{k-1}, r_{k-2}, \ldots, r_{k-n+1})]^{-\frac{1}{N}}, \qquad (4)$$

where N is the length of role sequence $R = \{r_1, \ldots, r_N\}$. The PP values are reported in Table 4, together with the ratio $PP/|\mathcal{R}|$ of the PP to the number of roles of each corpus.

The PP is the inverse of the geometric mean of $p(r_k|r_{k-1}, \ldots, r_{k-n+1})$ along a sequence R. Thus, when PP is low, this probability is, on average, high and roles from r_{k-n+1} to r_{k-1} influence significantly role r_k. The consequence is that only few roles can have probability significantly higher than 0 of appearing immediately after r_{k-1}. This corresponds to say that the roles are formal, that is the direct interaction (i.e., adjacency in R) between roles is more constrained.

Following the Kolmogorov-Smirnov Test [7], the difference between the performances achieved with HMMs and those achieved with the Bayesian classifier described in [4] are not

	C1		C2		C3	
	PP	$PP/\|\mathcal{R}\|$	PP	$PP/\|\mathcal{R}\|$	PP	$PP/\|\mathcal{R}\|$
1-gram	5.5	0.9	3.3	0.7	4.0	1.0
2-gram	2.1	0.4	2.5	0.5	3.0	0.8
3-gram	1.9	0.3	2.0	0.4	2.9	0.7

Table 4: PP stands for the perplexity measure of the different n-gram and $PP/|\mathcal{R}|$ is the proportion of the dictionary that has a probability higher than 0 to produce the n-gram sequence.

C1	HMM C	HMM W
Bayes C	77.1	2.7
Bayes W	5.4	14.8
C2	HMM C	HMM W
Bayes C	81.1	5.0
Bayes W	1.5	12.4
C3	HMM C	HMM W
Bayes C	27.1	14.0
Bayes W	11.1	47.7

Table 5: Diversity assessment. The table reports the percentage of data time where the two approaches are both correct (C), both wrong (W), or one wrong and the other correct.

statistically significant (see first columns of Table 2 and Table 3). However, the two classifiers show a significant degree of *diversity*, i.e. they make different decisions over the same sample in a relatively high percentage of cases (see Table 5). In particular, probabilistic sequential approaches tend to improve the recognition of less frequent roles that are typically penalized by Bayesian classifiers because of their low *a-priori* probability. *This suggests that the combination of the two approaches is likely to lead to significant performance improvements.* The highest possible performance deriving from a combination corresponds to the sum of the cases where at least one of the two approaches is right. This corresponds to 85.2% for C1, 87.6% for C2, and 52.2% for C3. In all of the cases, this would represent a statistically significant improvement with respect to the best of the approaches. The actual combination of the two approaches will be subject of future work.

4. CONCLUSIONS

This paper has presented an approach based on probabilistic sequential models (HMMs and n-gram language models) for the recognition of roles in different interaction settings. Two main findings result from the experiments: the first is that the Perplexity appears to be a good measure of how *formal* are the roles of a given setting, i.e. of how much they influence the interaction patterns of the people that play them. The second is that the comparison with the performance of a Bayesian classifier using the same features as this work shows that the two approaches are *diverse*, i.e. they make different decisions about the same sample in a significant fraction of cases.

The first finding is important because automatic recognition of roles is easier when these are formal [5], i.e. they are characterized by predictable, and machine detectable, behavioral patterns. To the best of our knowldge, this is the first work that proposes a quantitative measure of how formal roles are. The perplexity can be applied each time roles underly a sequence of events (speaker turns in the case of this work).

The second finding is important because it shows that the combination of two role recognition approaches promises to result into a signifcant improvement. This appears to be the case in particular for the meeting data (corpus C3), where the roles are *informal* (highest possible ratio $PP/|\mathcal{R}|$) and the two role recognizers have lower performance compared to the broadcast data (corpora C1 and C2).

The main open problem is the tuning of the D parameter (the number of events in the Social Affiliation Networks). In the case of meetings, fitting the D parameter to each recording (an approach that is not correct from a statistical point of view, but helps to understand the influence of the parameter) brings the accuracy to more than 65%. Thus, finding a better method to identify the best value for the D parameter will be the main subject of future work.

5. ACKNOWLEDGEMENTS

This work is supported by the Swiss National Science Foundation (under the NCCR on Interactive Multimodal Information Management), and by the European Community's Seventh Framework Programme (FP7/ 2007 – 2013), under grant agreement no. 231287 (SSPNet) and Petamedia. The authors wish to thank Hugues Salamin.

6. REFERENCES

[1] J. Ajmera and C. Wooters. A robust speaker clustering algorithm. In *Proc. of IEEE Workshop on Automatic Speech Recognition Understanding*, pages 411–416, 2003.

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

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

[4] S. Favre, H. Salamin, J. Dines, and A. Vinciarelli. Role recognition in multiparty recordings using social affiliation networks and discrete distributions. In *Proc. of International Conference on Mutlimodal Interfaces (ICMI)*, 2008.

[5] J. Levine and R. Moreland. Small groups. In D. Gilbert and G. Lindzey, editors, *The handbook of social psychology*, volume 2, pages 415–469. Oxford University Press, 1998.

[6] Y. Liu. Initial studay on automatic identification of speaker role in broadcast news speech. In *Proc. of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 81–84, 2006.

[7] F. Massey Jr. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, pages 68–78, 1951.

[8] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proc. of the IEEE*, volume 77, pages 257–286, 1989.

[9] R. Rosenfeld. Two decades of statistical language modeling: Where do we go from here. In *Proceedings of the IEEE*, volume 88, pages 1270–1278, 2000.

[10] H. Tischler. *Introduction to Sociology*. Harcourt Brace College Publishers, 1990.

[11] A. Vinciarelli, M. Pantic, and H. Bourlard. Social Signal Processing: Survey of an emerging domain. *Image and Vision Computing, to appear*, 2009.

[12] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.