Noname manuscript No.
(will be inserted by the editor)

Modeling the Synchrony Between Interacting People: Application to Role Recognition

Received: date / Accepted: date

Abstract The study of social interactions has attracted increasing attentions. The role recognition is one of its possible applications and the core of this study. This article proposes some approaches to automatically recognize the role of the participants of a meeting by modeling the synchrony of temporal nonverbal audio features. In our approache the Influence Model (IM), a Hidden Markov Model (HMM)-like, is used to model this synchrony and to extract from input data a feature vector that contains both information about temporal transitions (intra-personal data) and interaction between participants (inter-personal data). This modeling of the meeting is used as input of a Random Forests (RFs) for the role recognition task. The experiments are performed on 138 meetings (approximately 45 hours of recordings) from Augmented Multiparty Interaction (AMI) Corpus. Accuracy scores show that this combination of generative (IM) and discriminative (RFs) approaches permits to outperform state-of-the-art role recognition rates.

Keywords Role Recognition · Influence Model · Interaction Modelling · Synchrony

1 Introduction

The study of social intelligence becomes more and more popular, because of its importance for the analysis of social interactions. It was first defined by Edward Thorndike as "*the ability to understand and manage men and women, boys and girls, to act wisely in human relations*" [1]. The goal of social intelligence is then to understand the interaction between people by analyzing their exchanged social signals. One of the important tasks for the understanding of social interactions is the recognition of the role of each participant during the interaction, which has become an important application for social signal processing.

The role recognition has been addressed in two main contexts: during broadcast news and small scale meetings. Broadcast news involve strong interaction patterns and specific roles such as anchorman, journalist or interviewer. Although the recognition of such roles has become an important task for the understanding of global interactions, we are interested in this paper in the role recognition during small group interactions, *i.e.* on four-role based meeting scenarios, by mainly using speech features. This has many applications, for example for automatic indexing, retrieval and summarization of a meeting. This task requires, first to extract some relevant features from signals and to represent them in an appropriate way, then to learn some dominant patterns useful for the role recognition.

The features exchanged during social interactions can be classified into two main groups: verbal and nonverbal features. Verbal features contain lexical information and are highly dependent on the context: they are not easy to use for role recognition without any contextual information. A recent study showed that nonverbal features play a major role in the perception of social situations [2]. Moreover, there are two kinds of nonverbal features: global and temporal features. Global features, such as the total talking time of a participant, model the whole interaction and are often used by discriminative models like SVMs. On the opposite, temporal features capture temporal information, for example, the speech activity (who speaks at what time). They are usually used as input of generative models, like Hidden Markov Models or Influence Models, for the recognition. In this paper, we focus on temporal nonverbal features that appear during social interactions. To understand the underlying mechanisms of interactions, we propose to model the synchrony of these temporal features by using an Influence Model. Its output is used as the input of a discriminative method to automatically recognize the role of each participant of the meeting. To the best of our knowledge, this work is the first one to learn information provided by a dynamic model (*i.e* the Influence Model) using a static model (*i.e.* Random Forests).

The paper is organized as follows. Section 2 gives an overview of some works dedicated to the role recognition task. Section 3 presents the AMI database which is used in this paper to validate our approach, as well as the chosen methodologies. Section 4 then details the different configurations used for role recognition and presents our proposed approach. Comparative results are given in Section 5. In particular, we give technical details for all tested methodologies, then provide our results and compare them with other approaches' ones. Finally, conclusions and perspectives are given in Section 6.

2 Role recognition: state of the art

Many works have been done on role recognition. Some of them use global features, such as total talking time or total moving time, while others use

Table 1: Review of approaches dedicated to the role recognition task working on global or temporal features.

	Ref.	Features	Approach	Database	Acc.
Global	[3]	Total amount of speech of a participant	Decision Tree	Meetings (2 recordings, 5 roles)	53%
	[4]	Lexical and contextual features	HMM and Maximum Entropy method	TDT4 Mandarin broadcast news	77%
	[5]	Turn-taking and speaking time duration	Social Networks Analysis and Bayes classifier	Radio news bulletins (96 recordings, 6 roles)	85%
	[6]	Turn-taking and speaking time duration	Social Affiliation Network	AMI Corpus	56%
	[7]	Lexical, Turn-taking and speaking time duration	Social Affiliation Network	AMI Corpus	78%
	[8]	Multimodal nonverbal features	Social Networks Analysis	AMI Corpus (one role)	68%
	[9]	Event features	Time delay pattern	Music plays	N/A
	[10]	Nonverbal audio and video	Rule-based, Rank-level fusion, SVM and collective classification	ELEA	85.7%
	[11]	Nonverbal audio	SCP and IM	AMI Corpus	58%
Temp.	[12]	Speech activity and fidgeting	SVM	Mission Survival Corpus	65%
	[13, 14]	Speech activity and fidgeting	Influence Model	Mission Survival Corpus	75%
	[15]	Speech activity	Maximum likelihood	AMI Corpus	53%

temporal features such as speech activity (*i.e.* presence or absence of speech). We can then classify the role recognition approaches depending on the kinds of features (temporal and global) they work on. Note that, usually, temporal features are used by generative models, such as HMMs, that model temporal information, whereas global features are used by discriminative models, such as SVMs and Random Forests. However, in discriminative models, the temporal information, which is not taken into account during the classification step, can also be introduced into the features. Some related works are summarized in Table 1 and the two following subsections detail each family of approaches mainly working on speech features.

2.1 Approaches working on global features

To the best of our knowledge, the work in [3] is the first one on role recognition. It includes two main steps. First, the meeting is classified into two states: “Discussion” and “Information flow”. Then, four global features (number of times there is a change in speaker, total amount of speech of a participant, number of speech overlaps and average length of these overlaps) are used to recognize five roles (discussion participant, presenter, information provider, information consumer and undefined) using a decision tree. This approach was tested on recorded meetings based on the architecture proposed in [16] and reached an accuracy of 53%.

The work in [4] addresses the problem of role recognition (anchor, reporter and other) in broadcast news. The author uses two models for the classification: an Hidden Markov Model (HMM) [17] and a Maximum Entropy Model (MaxEnt), each one with different kinds of global features. Lexical and contextual features are used as observations in the HMM whose states are roles. For MaxEnt, three global features based on bigrams and trigrams are used for the classification. This approach was applied on the TDT4 Mandarin broadcast news database [4] and reached an accuracy of 81%.

The work in [5] aims at recognizing six roles (anchorman, second anchorman, guest, headline reader, weather man and interview participants) in broadcast news. Social Network Analysis (SNA) and Duration Distribution Modeling are used to extract the features from audio data (speaker segmentation). Role recognition is given by the posterior probability estimated using a Bayesian classifier. However, SNA requires a sufficient number of roles and planned scenarios to generate a meaningful interaction pattern. Furthermore, the role recognition is done independently for each person, without considering their interaction. This approach was tested on a database of radio news bulletins and reached and accuracy about 85%.

Salamin *et al.* [6] proposed to use Social Affiliation Network (SAN) to overcome the limitation of SNA used in [5]. Global features, extracted by SAN, including the information of duration of talk segments, are then used to estimate the posterior probability of each role. This approach was tested on both the radio news bulletins data [5] and on AMI meeting corpus [18]. Performances on the first dataset were slightly better than those in [5]. On the AMI meeting corpus, 56% of the total time was correctly labeled. Garg *et al.* [7] performed similar experiments and obtained an overall accuracy of 50%. They improved the accuracy to 78% by using features including lexical information.

Jayagopi *et al.* [8] have also tested their approach on AMI meetings. They first extract global audio features (such as intervention length or number of talk turn transitions between participants) and global video features (such as the total quantity of movement). The role of project manager as well as the most dominant participant are estimated. The project manager is recognized using the centrality measure [19] and the accuracy reached 68.4%.

Varni *et al.* detect the leader over a small group of music players in [9]. First, they extract event features, *i.e.* when does an event begin or end. Then they measure the relative time delay patterns between events for each pair of participants in the whole group. Finally, they recognize from these patterns the leader according to two metrics, leader rank and leader sum. There is no precise classification accuracy mentioned in their paper, but their results have been validated by professional music players. Note that although this approach uses global features, it also uses recurrence matrices to describe interactions between temporal features.

Emergent leaders are identified in [10] on the Emergent LEAder corpus [20]. Global features, such as total speaking time length, head activity time length or body activity time length are extracted from audio and video signals. Then four approaches are evaluated and compared for the recognition. The first one, a rule-based approach, considers that the participant with the longest total speaking time is the emergent leader. The second approach, a rank-level approach, determines who is the emergent leader by ranking features. The third approach combines all features into a single vector given as input of a SVM to detect the emergent leader. The last one, a collective classification approach, uses an Iterative Classification Algorithm [21] for the classification. Finally, this last approach reached the best accuracy of 85.7%.

Finally, Cristani *et al.* use a coding method named Steady Conversational Period (SCP) and IM to recognize the role of each participant in AMI corpus [11]. SCPs are used to segment the whole meeting into steady periods. A Gausian Mixture Model is then used to cluster these periods into long silence, short silence, long speech or short speech. After the clustering, IM is used with the SCP for the recognition. This approach gets an overall accuracy of 58%.

2.2 Approaches working on temporal features

Temporal features, also called local features, contain some dynamical information. Because of the huge number of temporal features that can be extracted from data, the classification process suffers from the curse of dimension and over-fitting problems, that some approaches have tried to solve.

The approach in [12] recognizes the roles on the database Mission Survival Corpus [22]. It uses a SVM to process two local features. Original audio and video data sequences are divided into small time-windows (*i.e.* 10 seconds/window), from which local features (speech activity and fidgeting) are extracted. These extracted features are concatenated into a vector which is the input of a SVM. Two experiments are conducted: task role recognition (follower, orienter, giver, seeker and recorder) and socio-emotional role recognition (neutral, gate-keeper, supporter, protagonist and attacker). This approach got an accuracy of 65% for the first experiment and of 70% for the second one. However, although SVMs are robust to the over-fitting problem, they still suffer from the curse of dimension.

The approaches in [14,13] solve the problems of the curse of dimension and over-fitting by using the Influence Model (IM) [23]. This model significantly reduces the number of needed parameters and considers the influence between participants of the meeting (see Section 3.2 for more details). Temporal features speech activity and fidgeting are used directly as input of the IM. The likelihood given by IM is used for the role recognition. The performance of this approach on the Mission Survival Corpus database got an overall accuracy of 63% on Task Area and Socio-Emotional Area systems.

Similarly to the approaches in [14,13], the work in [15] extracts speech activities, named talk-spurts, whose interval pause length is lower than 0.3 seconds. Then four probabilities are computed according to these talk-spurts: the probability that a participant starts speaking at time t when no-one else was speaking at t-1, the probability that a participant continues speaking at time t when no-one else was speaking at t-1, the probability that a participant starts speaking at time t when another participant was speaking at t-1 and the probability that a participant continues speaking at time t when another participant was speaking at t-1. According to these probabilities, a maximum likelihood criteria is used to recognize the roles. This approach was tested on the AMI database for a four-role recognition task and reached an accuracy of 53%.

3 Data and methodologies

This paper proposes a formalism to model several signals with temporal dependencies and to use this model of synchrony in a role recognition task. Before introducing our approach, we present the database AMI and the main methodologies that have been proposed in the literature to model interdependent signals.

3.1 Extracted data: nonverbal audio features of AMI Corpus

The database AMI Corpus has been created by the European-funded AMI project and is online available (`http://groups.inf.ed.ac.uk/ami/corpus/`). It includes about 100 hours of recorded meetings. Each recording contains several signals which are synchronized to a common time-line. These signals include close-talking and far-field microphones, individual and room-view video cameras, output from a slide projector and from an electronic white-board. During the meetings, the participants also use unsynchronized pens to record their writing. There are two kinds of meetings, scenario meetings and non-scenario meetings. All the approaches presented below have been tested on the 138 scenario meetings in which participants play the roles of employees in an electronics company that decide to develop a new type of television remote control. Each participant has a specific role to play: the project manager (PM), the marketing expert (ME), the user interface designer (UI) and the industrial designer (ID).

We only consider the speech activities of participants as features and, more specifically, the binary representation of speaking activities, *i.e.* 1: speak, 0: do not speak. This information was manually extracted from speech transcriptions provided by AMI Corpus Database, but this step could also be easily automatized by thresholding speech energy. Except explicitly specified, the input data is a binary signal for each participant with a sampling rate of 4 Hz.The signal is represented by a series of values for each participant (see Section 4). These signals and their dependencies (synchrony or interaction) can be modelded using Hidden Markov Models, Coupled Hidden Markov Models or Influence Models.

3.2 Hidden Markov Model and Influence Model

Suppose the behavior of one person is modeled by a Markov chain with Q distinct state values $\{s^1, \ldots, s^Q\}$. In such a case, the transition matrix of size $Q \times Q$ is composed of elements $P(S_t|S_{t-1})$, where S_t is the hidden state at time t taking value in $\{s^1, \ldots, s^Q\}$.

The problem now is to model the interaction between C interacting people. A first solution is to consider a single chain with Q^C state values, that leads to a transition matrix with $Q^C \times Q^C$ elements. The problem quickly becomes intractable when C increases. Another solution is to use the Coupled Hidden Markov Models (CHMM), first introduced by Brand *et al.* [24]. The structure of a CHMM with C chains leads to $C \times Q$ state values $s^{i,k}$, with $i = 1, \ldots, Q$ and $k = 1, \ldots, C$. Here, S_t^k, the hidden state of chain k at time t, depends on the previous state of all the chains at time $t-1$. Transition probabilities are then given by $P(S_t^k|S_{t-1}^1, S_{t-1}^2, \ldots, S_{t-1}^C)$, with $k = 1, \ldots, C$. The transition matrix now contains $C \times Q^{C+1}$ elements, which also becomes intractable when C increases.

The Influence Model (IM) [25] keeps the same graphical model as CHMM, but introduces the following simplification:

$$P(S_t^k|S_{t-1}^1, S_{t-1}^2, \ldots, S_{t-1}^C) = \sum_{l=1}^{C} t^{l,k} P(S_t^k|S_{t-1}^l) \tag{1}$$

where $t^{l,k}$ denotes the influence of chain l on chain k. Thus, the transition probabilities are now expressed as a linear combination of pairwise conditional probabilities. We have now $(C \times Q)^2$ elements $P(S_t^k|S_{t-1}^l)$ and C^2 parameters $t^{l,k}$. This gives a total number of $(C \times Q)^2 + C^2$ parameters, that is lower than those of the two previous models. Another advantage of IM is that parameters $t^{l,k}$ have a direct interpretation in the context of interaction. In Figure 1, we give an example of interactions between three participants k, l and m, where each one is modeled by a Markov chain.

Formally, an IM for C chains is represented by the set of parameters $\lambda = \{S, \pi, A, T\}$. S is a vector containing the states which can be hidden or observed, π is a vector containing the initial probabilities of each state of S, A is the matrix containing conditional probabilities $P(S_t^k|S_{t-1}^l)$ and T is the matrix containing influence coefficients $t^{l,k}$ between participants l and k, with $k, l = 1, \ldots, C$.

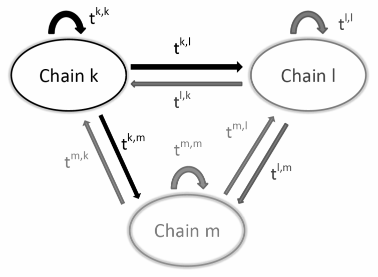


Fig. 1: The Influence Model of an interaction between three participants k, l and m.

Matrix A contains $C \times C$ sub-matrices $A^{l,k}$ of size is $Q \times Q$. Sub-matrices $A^{l,k}$ represent the conditional probabilities of states in chain k during its interaction with chain l. More generally, if each chain i has n_i state values, then the dimension of each $A^{l,k}$ is $n_l \times n_k$ and the total dimension of A is $\sum_{i=1}^{C} \sum_{j=1}^{C} n_i n_j$.

Influence matrix T contains the influence coefficients $t^{l,k}$, which represent the influence of chain l on chain k. The dimension of T is then $C \times C$. A new matrix H is derived from matrices A and T and is composed of $C \times C$ sub-matrices $H^{l,k}$ such as:

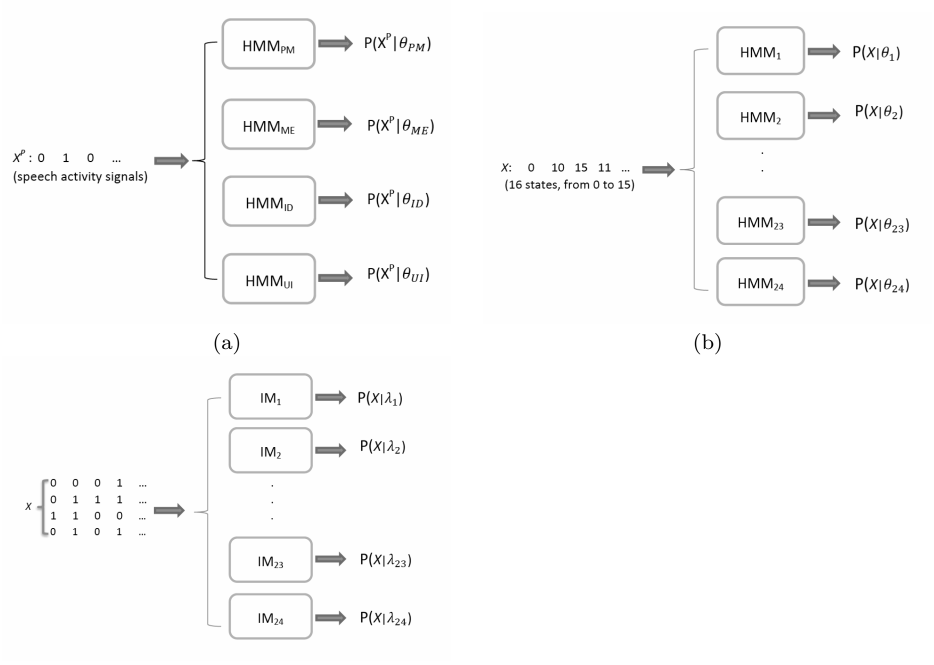
$$H^{l,k} = A^{l,k} \times t^{l,k}$$

This new matrix contains both the information of transition probability of matrix A and of influence of matrix T. In our context of interaction modeling, H characterizes the interaction and is used as feature vector.

4 Role recognition

As specified in Table 1, the synchrony can be modeled using global or temporal features and the role recognition can be done using either generative or discriminative models. HMMs and IMs are well known to model temporal and interdependent signals. But as said by Bishop *et al.* [26] and confirmed by Ng *et al.* [27], *“the generalization performance of generative models is often found to be poorer than that of discriminative models due to differences between the model and the true distribution of the data”*. It is the reason why several authors [28,29,30,31] combine the strengths of both models.

In this section we detail how to recognize role by using generative approaches (HMMs and IMs). We also present our proposed approach consisting in combining a generative and a discriminative approach for the role recognition task.



(a) (b)

(c)

Fig. 2: Classification approaches based on HMM and IM: (a) One HMM for each role (HMM4R), (b) one HMM for all the participants (HMM4M) and (c) an IM for the whole meeting (IM)

4.1 Role recognition using one HMM for each role (HMM4R)

In the first method, named HMM4R and illustrated in Figure 2(a), one HMM, with 2 observed states, has been trained by the binary representation of speaking activity of only one participant. We then have the set $\mathrm{HMMs} = \{\mathrm{HMM_{PM}}, \mathrm{HMM_{ID}}, \mathrm{HMM_{ME}}, \mathrm{HMM_{UI}}\}$ and corresponding parameter sets $\{\vartheta_{\mathrm{PM}}, \vartheta_{\mathrm{ID}}, \vartheta_{\mathrm{ME}}, \vartheta_{\mathrm{UI}}\}$ (each parameter set only contains parameters S, π and A from the set λ previously defined for IM). For the role recognition, the input data $\mathrm{X^p}$ provided by the speech activity of participant p is evaluated by each of the four HMMs (for each one, π and A have been randomly initialized). By applying Forward-Backward algorithm, we get $\mathrm{P(X^p|\vartheta_r)}$, where $\mathrm{P(X^p|\vartheta_r)}$ is the likelihood generated by $\mathrm{X^p}$ and $\mathrm{HMM_r}$, $\mathrm{r} \in \{\mathrm{PM}, \mathrm{ID}, \mathrm{ME}, \mathrm{UI}\}$. Then, participant p is associated to the role $\mathrm{\hat{r}_p}$ whose HMM gave the highest likelihood, *i.e.* such that:

$$\mathrm{\hat{r}_p} = \arg\max_{\mathrm{r}} \mathrm{P(X^p|\vartheta_r)} \tag{2}$$

4.2 Role recognition using one HMM (HMM4M) for the whole meeting

The second method, called HMM4M, models the whole interaction between the four people with one conventional HMM. At each time step, a decimal number, converted from four binary numbers, is used to represent the meeting states (for example, 0=0000 means nobody speaks, 8=1000 means only the first participant speaks, 12=1100 means participants 1 and 2 are speaking, *etc.*). We have now 16 possible observed values (from 0 to 15). The input data is now a series X of these observed values.

For the AMI meeting with 4 roles, there are $A_4^4 = 24$ possible configuration orders for the 4 roles, as shown in Table 2 (one configuration per column). A HMM (whose states correspond to the 16 observed values) is trained for each one of the 24 configurations (see Figure 2(b)) to get parameter sets $\vartheta_1, \vartheta_2, \ldots, \vartheta_{24}$. Note that, during the training process, some "artificial" data are generated by re-ordering original data according to the 24 different configurations for each meeting. As for previous approaches, π and A have been randomly initialized for each HMM.

By applying Forward-Backward algorithm, we get $P(X|\vartheta_r)$, where $P(X|\vartheta_r)$ is the likelihood generated by X and HMM_r, $r \in \{1, \ldots, 24\}$. The role classification is made by selecting the HMM (among the 24 possible), which gives the highest likelihood for the input data, to get the role configuration

$$\hat{r} = \arg\max_r P(X|\vartheta_r) \tag{3}$$

Table 2: 24 order configurations (one per column) for 4 roles

PM	PM	PM	PM	PM	ID	...
ID	ID	ME	ME	ID	PM	...
ME	UI	UI	ID	UI	ME	...
UI	ME	ID	UI	ME	UI	...

4.3 Role recognition using Influence model (IM)

In the third method, called IM, an Influence Model is used to model the whole meeting. The training and classification processes are the same as for the approach HMM4M, described in Section 4.2. However, the input data X of IM are the four observed binary chains, one for each participant (series of 0 and 1 as in Section 4.1). We then have a total number of 8 observed values. Similarly to the previous approach, 24 IM are learned, one for each role configuration (see Figure 2(c)), that gives parameters sets $\{\lambda_1, \lambda_2, \ldots, \lambda_{24}\}$. We get $P(X|\lambda_r)$, where $P(X|\lambda_r)$ is the likelihood generated X and IM_r, $r \in \{1, \ldots, 24\}$. The IM getting the highest likelihood gives the role configuration:

$$\hat{r} = \arg\max_r P(X|\lambda_r) \tag{4}$$

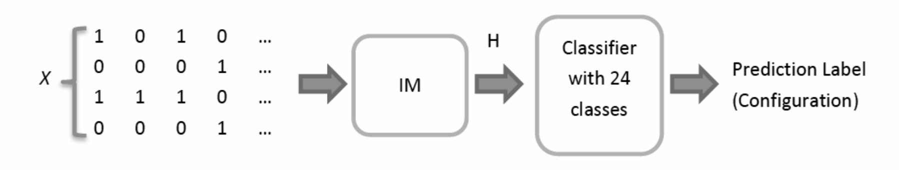


Fig. 3: The approach using IM and discriminative classifier with 24 classes

It is difficult to improve the performance of generative approaches, because they suffer from their low power in discrimination. One solution is to extract global features from sequences and then to use discriminative classification methods, such as the approach proposed in [10]. But in this case, we face the problem of choosing relevant global features. Moreover, it is difficult to introduce concepts such as synchrony [32] into these global features. Thus, we propose to, first use a generative model to describe the data, then to use a discriminative method for the classification as presented in the following subsection.

4.4 Proposed approach: role recognition by mixing generative and discriminative methods

As in the previous subsection, each meeting is modeled by an IM with 4 chains, one chain per participant and 2 state values per chain. This modeling generates a matrix H that contains 4×4 sub-matrices of size 2×2. H includes both information about transition probability and influence between chains. We transform H into a feature vector of size 64 for the classification. Then two solutions are used to recognize the four roles.

In the first solution, illustrated in Figure 3, a single classification is made with 24 labels corresponding to the 24 role configurations of a meeting as presented in Table 2. The second solution, illustrated in Figure 4, consists in 4 independent classifications (one for each role) with 4 labels (the position of the role in the meeting). In this work, the classification is made by RFs, while SVM could also be used. This leads to two methods called IMRF4 and IMRF24.

5 Comparative results

We present in this section the different methods that have been studied for role recognition and detailed in Section 4. As specified in Section 3.1, only the speech activity of participants is considered and more specifically, the binary representation of speaking activities (*i.e.* 1: speak, 0: do not speak). The global features have been extracted on the whole meetings. Using these data, we recall we compare five role-classification methods (see Section 4 for details):

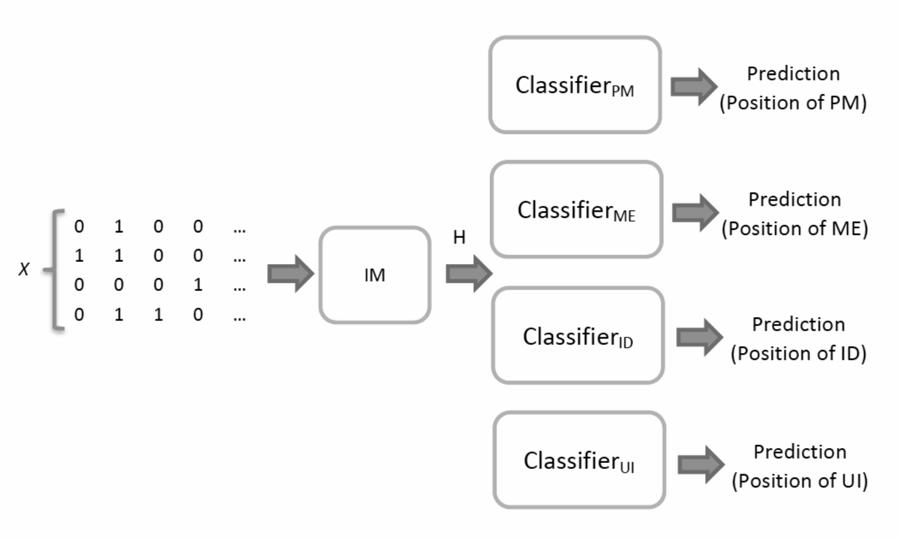


Fig. 4: The approach using IM and discriminative classifier with 4 classes

Table 3: Accuracies of tested and existing methods for the role recognition task on AMI database

Method	PM	ID	ME	UI	Overall
HMM4P	0.7	0.05	0.1	0.85	0.42
HMM4M	0.85	0.3	0.4	0.55	0.52
IM	0.75	0.25	0.46	0.49	0.49
IMRF24 CV	0.8	0.4	0.54	0.44	0.55
IMRF24 VS	0.76	0.42	0.52	0.45	0.54
IMRF4 CV	0.71	0.56	0.56	0.49	**0.58**
IMRF4 VS	0.75	0.47	0.53	0.47	0.56
[6]	0.76	0.41	0.38	0.41	0.56
[7]	0.79	0.25	0.20	0.45	0.50
[8]	0.65	/	/	/	/
[15]	0.6	0.4	0.4	0.7	0.53
[11]	0.85	0.4	0.6	0.5	0.58

- One HMM for each role (HMM4R).
- One global HMM (HMM4M).
- Influence model (IM).
- Influence Model and RF with 24 classes (IMRF24).
- Influence Model and RF with 4 classes (IMRF4).

Furthermore, AMI database has been divided into three subsets by the authors: 98 meetings for training, 20 meetings for validation and 20 meetings for testing. We kept this protocol for all approaches, except when explicitly specified.

5.1 One HMM for each role (HMM4R)

The accuracy scores obtained with this method are given in the first row of Table 3. One can see that some roles are well recognized (PM, UI), while others have a very low recognition rate. The main drawback of this approach is that the decisions are made individually without considering all the participants. Several participants can then be associated to the same role during the meeting. That is why some roles (ID and ME) have a very low accuracy. Thus, no interacting information is considered in this model. A first improvement is to consider simultaneously all participants and to model the whole meeting with one HMM.

5.2 One global HMM (HMM4M)

Role recognition accuracies are given in the second row of Table 3. The global role recognition accuracy (0.52) is higher than for the previous approach HMM4R (0.42). This proves the interest of considering all participants of the meetings rather than only the participant whose role is to be predicted. However, even if there are only 16 state values, for this case with 4 participants, a model that could describe the behavior of each participant as well as the dependences between them should be more suitable.

5.3 Influence model (IM)

The role recognition accuracies are given in the third row of Table 3. The accuracies are slightly lower than those for approach HMM4M. In this application, the small number of state values required by IM does not justify the interest of the influence model even if it introduces some simplifications in the formalism as explained in subsection 3.2. Actually, with only 16 state values in the HMM4M approach, the problem remains tractable with a single chain. As IM is a simplified version of HMM4M, it gives an approximation of the exact accuracies provided by HMM4M, that explains why it gets lower accuracies. Nevertheless, a solution using IM can deal with scenarios of meetings involving more participants. For example, for $\mathrm{C} = 16$ interacting participants and $\mathrm{Q} = 2$ states per participant, the number of elements of the transition matrix is $(\mathrm{C} \times \mathrm{Q})^2 = 1024$ for IM and $\mathrm{Q}^{2\mathrm{C}} \simeq 4.3 \times 10^9$ for HMM4M.

5.4 Generative modeling and discriminative classification (IMRF24 and IMRF4)

The parameters of RFs have been optimized either with the validation set proposed by AMI (VS) or using a K-folder cross-validation (CV): the training set is divided into K parts, $\mathrm{K} - 1$ are used for training and the last one for validation. Among the K parameter sets ($\mathrm{K} = 5$ in this application), the one with the best average performance is used in the testing set. All these results are given in rows 4-7 of Table 3. As can be seen, these two new approaches outperform the previous ones.

5.5 Influence of the temporal scale

In this section, we study the influence of the temporal scale on method IMRF4. Actually, this scale is important since IM is based on the study of the transition between consecutive state values. As stated before, the frequency of audio data is 4 Hz (a sampling period of 0.25 seconds). Data are rescaled at frequencies 10 Hz (0.1 s.), 1 Hz (1 s.), 0.2 Hz (5 s.) and 0.1 Hz (10 s.). Table 4 gives the accuracies obtained by IMRF4 for each of these scales.

Table 4: Accuracies for the role recognition task given by method IMRF4 depending on the temporal scale in Hz.

Scale	Data Set	PM	ID	ME	UI	Overall
10 Hz	CV	0.63	0.21	0.38	0.22	0.36
4 Hz	CV	0.71	0.56	0.56	0.49	**0.58**
1 Hz	CV	0.76	0.36	0.43	0.45	0.50
0.5 Hz	CV	0.74	0.45	0.27	0.35	0.45
0.1 Hz	CV	0.84	0.33	0.18	0.32	0.42

The frequency of data influences the value of the elements in matrix H. As can be seen in Table 4, the optimum sampling period, to not consider silences between words for example, is 4 Hz (0.25 s.). A smaller period would introduce silences between words that are not informative, while a higher one would delete small words. The confusion matrix obtained with this frequency is given in Table 5. In order to get stable results, we have generated 480 testing samples from the 20 testing meetings by rearranging the 4 roles into 24 different orders, as introduced in Section 5.2, while the experiments in [11] only consider 20 testing samples. While the project manager is well recognized, there is a confusion between the other roles.

Table 5: Confusion matrix for IMRF4 with a 4Hz sampling rate.

	PM	ID	ME	UI
PM	342	58	48	32
ID	28	267	68	117
ME	14	81	269	116
UI	50	109	84	237

5.6 Comparative study

The lower part of Table 3 gives accuracies for the four-role recognition problem obtained by other works on AMI database.

Our accuracies are higher than the best of the first four approaches in lower part. However, we must be careful on this comparison because the role prediction results are not estimated in the same way. We estimate the role of each participant by studying the whole meeting, while in [6] the role of the speaker is only estimated on a short temporal window. Works in [11] get the same overall recognition accuracy as our. However, there are some differences between the two approaches and experiments. Our approach focusses on temporal features extracted at each time step, while they use steady periods as time steps with attribute "long" or "short", (*i.e.* long speech, short silence). Thus, this coding makes a strong hypothesis concerning conversations, as it only considers two possible durations for steady conversations. On the contrary, our approach does not require any coding as pre-processing step: this coding is implicitly done by the IM. Actually, the duration is estimated by the transition probability.

Back to the comparison of the five models introduced in this paper, several conclusions can be made. First, even if the goal is to estimate the role of each person, it is better to consider all the participants (HMM4M - 0.52) rather than just the one we want to estimate the role (HMM4P - 0.42). For this modeling of the four participants, two generative methods were tested: HMM4M and IM. The results are roughly comparable (0.52 and 0.49 respectively), however, the influence model remains tractable when the number of participants or the number of state values per chain increases, while it is not the case for HMM as explained in Section 5.3. Finally, the two last approaches, IMRF24 and IMRF4, that combine the strengths of generative and discriminative models, lead to the best recognition rates, respectively 0.55 and 0.58.

6 Conclusion

In this article, we have presented a new approach for role recognition in small scale meetings that has been validated on AMI database. Our approach models the synchrony of nonverbal audio features of all participants by using an Influence Model. The output of this model is then used as input feature of a Random Forests classifier for the role recognition task. This has three main advantages. First, the complexity is greatly decreased by using IM compared to using HMM or coupled HMM, particularly when the number of participants becomes high, or when the number of state values increases. Secondly, the matrix generated by an Influence Model encodes both intra and inter personnel information, so it is a good descriptor of the interaction between a group of people. Finally, compared to using a single IM scheme for role classification, applying a combination of generative (IM) and discriminative (RFs) models greatly improves the recognition rates (from 0.49 to 0.58 on AMI database).

Our results outperform those obtained with the SAN approach [6], the best up-to-date performances on database AMI. Our current works focus on integrating multi-modal information into our recognition model. Indeed, adding information such as quantity of movement or gaze should improve the classification rates.

References

1. E. Thorndike, "Intelligence and its use," *Harper's Magazine*, vol. 140, pp. 227–235, 1920.
2. A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.
3. S. Banerjee and A. I. Rudnicky, "Using simple speech–based features to detect the state of a meeting and the roles of the meeting participants," in *International Conference on Spoken Language Processing*, pp. 1–4, 2004.
4. Y. Liu, "Initial study on automatic identification of speaker role in broadcast news speech," in *Conference of the North American Chapter of the Association for Computational Linguistics, Human Language Technology*, pp. 81–84, 2006.
5. A. Vinciarelli, "Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling," *IEEE Transactions on Multimedia*, vol. 9, no. 6, pp. 1215–1226, 2007.
6. H. Salamin, S. Favre, and A. Vinciarelli, "Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1373–1380, 2009.
7. N. P. 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 *MM*, pp. 693–696, 2008.
8. D. B. Jayagopi, S. Ba, J.-M. Odobez, and D. Gatica-Perez, "Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues," in *International Conference on Multimedia Interaction*, pp. 45–52, 2008.
9. G. Varni, G. Volpe, and A. Camurri, "A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 576–590, 2010.
10. D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez, "A nonverbal behavior approach to identify emergent leaders in small groups," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 816–832, 2012.
11. M. Cristani, A. Pesarin, C. Drioli, A. Tavano, A. Perina, and V. Murino, "Generative modeling and classification of dialogs by a low-level turn-taking feature," *Pattern Recognition*, vol. 44, no. 8, pp. 1785–1800, 2011.
12. M. Zancanaro, B. Lepri, and F. Pianesi, "Automatic detection of group functional roles in face to face interactions," in *International Conference on Multimedia Interaction*, pp. 28–34, 2006.
13. W. Dong, B. Lepri, F. Pianesi, and A. Pentland, "Modeling functional roles dynamics in small group interactions," *IEEE Transactions on Multimedia*, vol. 15, no. 1, pp. 83–95, 2013.
14. W. Dong, B. Lepri, A. Cappelletti, A. S. Pentland, F. Pianesi, and M. Zancanaro, "Using the influence model to recognize functional roles in meetings," in *International Conference on Multimedia Interaction*, pp. 271–278, 2007.
15. K. Laskowski, M. Ostendorf, and T. Schultz, "Modeling vocal interaction for text-independent participant characterization in multi-party conversation," in *Workshop of Special Interest Group on Discourse and Dialogue*, pp. 148–155, 2008.
16. S. Banerjee, J. Cohen, T. Quisel, A. Chan, Y. Patodia, Z. Al Bawab, R. Zhang, A. Black, R. M. Stern, A. I. Rudnicky, *et al.*, "Creating multi-modal, user-centric records of meetings with the carnegie mellon meeting recorder architecture," in *International Conference on Acoustic, Speech and Signal Processing*, 2004.

17. L. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
18. I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, *et al.*, “The AMI meeting corpus,” in *Measuring Behavior*, vol. 88, 2005.
19. S. Wasserman, *Social network analysis: Methods and applications*, vol. 8. Cambridge University Press, 1994.
20. D. Sanchez-Cortes, O. Aran, and D. Gatica-Perez, “An audio visual corpus for emergent leader analysis,” in *Multimodal Corpora*, 2011.
21. L. K. McDowell, K. M. Gupta, and D. W. Aha, “Cautious collective classification,” *The Journal of Machine Learning Research*, vol. 10, pp. 2777–2836, 2009.
22. F. Pianesi, M. Zancanaro, B. Lepri, and A. Cappelletti, “A multimodal annotated corpus of consensus decision making meetings,” *Language Resources and Evaluation*, vol. 41, no. 3-4, pp. 409–429, 2007.
23. C. Asavathiratham, *The influence model: A tractable representation for the dynamics of networked Markov chains.* PhD thesis, MIT, 2000.
24. M. Brand, N. Oliver, and A. Pentland, “Coupled hidden Markov models for complex action recognition,” in *Computer Vision and Pattern Recognition*, pp. 994–999, 1997.
25. S. Basu, T. Choudhury, B. Clarkson, A. Pentland, *et al.*, “Learning human interactions with the influence model,” in *Conference on Neural Information Processing Systems*, 2001.
26. J. Lassere and C. Bishop, “Generative or discriminative? getting the best of both worlds,” *Bayesian Statistics*, vol. 8, pp. 3–24, 2007.
27. A. Ng and M. Jordan, “On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes,” *Conference on Neural Information Processing Systems*, vol. 14, p. 841, 2002.
28. A. Holub and P. Perona, “A discriminative framework for modelling object classes,” in *Computer Vision and Pattern Recognition*, pp. 664–671, 2005.
29. J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, “Generative or discriminative? getting the best of both worlds,” *Bayesian Statistics*, vol. 8, pp. 3–24, 2007.
30. M. Salzmann and R. Urtasun, “Combining discriminative and generative methods for 3d deformable surface and articulated pose reconstruction,” in *Computer Vision and Pattern Recognition*, pp. 647–654, 2010.
31. R. Rosales and S. Sclaroff, “Combining generative and discriminative models in a framework for articulated pose estimation,” *International Journal of Computer Vision*, vol. 67, no. 3, pp. 251–276, 2006.
32. E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen, “Interpersonal synchrony: A survey of evaluation methods across disciplines,” *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 349–365, 2012.