Pattern Recognition 44 (2011) 1785–1800

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Generative modeling and classification of dialogs by a low-level turn-taking feature

Marco Cristani [c,*,1], Anna Pesarin [a], Carlo Drioli [a], Alessandro Tavano [b], Alessandro Perina [d], Vittorio Murino [c,*,1,2]

[a] Dipartimento di Informatica, University of Verona, Strada le Grazie 15, 37134 Verona, Italy
[b] Institute of Psychology, University of Leipzig, Seeburgstr. 14-20, 04103 Leipzig, Germany
[c] Istituto Italiano di Tecnologia (IIT), Via Morego 30, 16163 Genova, Italy
[d] Microsoft Research, One Microosoft Way, 98052 Redmond, WA

ARTICLE INFO

Article history:
Received 22 March 2010
Received in revised form
4 January 2011
Accepted 19 January 2011
Available online 27 January 2011

Keywords:
Dialog analysis
Generative modeling
Classification
Feature extraction

ABSTRACT

In the last few years, a growing attention has been paid to the problem of human–human communication, trying to devise artificial systems able to mediate a conversational setting between two or more people. In this paper, we propose an automatic system based on a generative structure able to classify dialog scenarios. The generative model is composed by integrating a Gaussian mixture model and a (observed) Markovian influence model, and it is fed with a novel low-level acoustic feature termed steady conversational period (SCP). SCPs are built on duration of continuous slots of silence or speech, taking also into account conversational turn-taking. The interactional dynamics built upon the transitions among SCPs provides a behavioral blueprint of conversational settings without relying on segmental or continuous phonetic features, and may be important for predicting the evolution of typical conversational situations in different dialog scenarios. The model has been tested on an extensive set of real, dyadic and multi-person conversational settings, including a recent dyadic dataset and the AMI meeting corpus. Comparative tests are made using conventional acoustic features and classification methods, showing that the proposed scheme provides superior classification performances for all conversational settings in our datasets. Moreover, we prove that our approach is able to characterize the nature of multi-person conversation (namely, the role of the participants) in a very accurate way, thus demonstrating great versatility.

1. Introduction

In the recent years, there has been a growing interest in the development of the so-called conversation "external mediators", *i.e.*, dialog analysis systems that observe human beings interacting with each other and assist them by capturing their conversational behavior. Such computer devices in the human interaction loop would possibly provide feedback to enhance the human–human communication and relations in general [1–7], concurring to the development of the so-called *human computing* scientific area [8]. An important aim of an external mediator is to obtain a good yet general blueprint of a dialog situation by analyzing the ongoing conversational dynamics, intended as the alternating speech behavior exploited by the partners during negotiation [2,4,3]. The ability to carefully capture and classify conversational dynamics could also be employed to enhance the performance of a wide range of applications, such as dialog detection [9], speaker recognition/verification [5], and event detection in meeting scenarios [10], also considering video cues. More consistently, it would improve social signalling applications [1–3,11–14], such as the ones that link conversational dynamics to social roles (*e.g.*, *dominance* [15], *mirroring* [3] and others [1]), or those that face interesting and complex challenges such as the "thin slice" detection, *i.e.*, the ability of predicting the outcome of a specific conversational exchange in very limited time [13].

In this paper, we present a dialog analysis system which provides a statistical signature, *i.e.*, a set of model parameters, characterizing different audio profiles among various conversational situations, proposing a novel way to *encode* and *explain* the conversational dynamics.

The key characteristic of our approach is represented by a serial generative framework, composed by a Gaussian mixture model (GMM) [16] followed by an observed influence model [3]

* Corresponding authors. Tel.: +39 45 8027988; fax: +39 45 8027068.
E-mail addresses: marco.cristani@univr.it (M. Cristani), anna.pesarin@univr.it (A. Pesarin), carlo.drioli@univr.it (C. Drioli), tavano@uni-leipzig.de (A. Tavano), alessandro.perina@gmail.com (A. Perina), vittorio.murino@univr.it (V. Murino).
[1] The authors are also with the Dipartimento di Informatica, University of Verona, Italy.
[2] Tel.: +39 45 8027996; fax: +39 45 8027068.

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.01.013

at the top level. Such framework is fed by a novel type of simple, low-level auditory features, which are termed *steady conversational periods* (SCPs). These are built on duration of continuous slots of silence or speech, also taking into account conversational turn-taking, so allowing to easily capture and profile the silence/speech dependencies in dialogs.

This special combination of features and related (generative) processing constitutes the actual contribution of the work, aiming not only at classifying dialogs, but also at *explaining* the nature of the dialogs, for example characterizing the role of the participants.

Despite their simplicity, the proposed system is able to analyze difficult conversational situations. In the experiments, we show that very high accuracy is reached in discriminating conversations in which a preschool child and an adult are involved. Further, we present results that express the capability of identifying conversational mood, *e.g.*, understanding whether two persons are discussing normally or arguing. At the best of our knowledge, this is the first time that such situations are processed in an automatic way. Beside classifying different conversational settings, the system is able to finely characterize specific conversational behaviors owing to the generative approach employed, which mirrors subtle behavioral aspects through different model parameterizations. More specifically, it is actually able to capture the attitude of self-selecting for turn-taking even though the interlocutor has not yet completed his own turn; further, it indirectly models speech planning by characterizing the tendency to utter short sentences instead of longer propositions.

In the experiments, we also show how these aspects can be exploited to model each single dialog participant, classifying its role in a meeting. This promotes the generative use of SCPs in a wide range of applications.

The use of the SCP, *i.e.*, focusing on the alternation between speech and silent segments both within and between subjects, can be well motivated from a behavioral viewpoint. At a physiological level, the timed coordination of speech activity between two participants is constrained by the respiratory kinematics of preparation to speech initiation (within subjects) and adaptation to initiated speech (between subjects, turn-taking skills) [17]. From a neurophysiological viewpoint, conversational timing skills are likely to be primarily sustained by a network of brain structures, among which mirror neurons and prefrontal cortex, allowing for embodied perspective-sharing among interlocutors [18,19]. Interestingly, recent neuropsychological research suggested that communicative perspective-sharing is supported by frontal neural networks that determine the timed inhibition of a subject's own perspective [20]. Thus, it becomes feasible to consider the conversational interplay of speech and silence as the joint product of intention understanding and self-monitoring.

In summary, this paper aims at presenting a brand new classification framework able to take into account a novel low-level auditory feature related to speech/silence alternation in a dialog in order to effectively model conversational situations.[3]

An extensive set of comparative experiments on real multi-person conversational data has been performed to test the proposed modeling architecture.

The rest of the paper is organized as follows. Section 2 summarizes the most related literature. In Section 3, the relevant mathematical background is provided, and specific details of the proposed framework are illustrated in Section 4. In Section 5, experimental results are reported using an extensive set of dialog conversational settings, and, finally, in Section 6 conclusions are drawn and perspective applications are envisaged.

[3] We presented a shorter and preliminary version of the present work in [21].

2. Related work

The automatic analysis of the speech signals for dialog modeling is usually performed with the aim of providing "some" information required to characterize a conversation. This processing can achieve various levels of complexity and can be driven by different motivations and goals, hence, the related literature is huge and multifaceted. To mention some of the most relevant analysis approaches, we can recall the most studied which include (1) dialog acts classification, which refers to the identification of the nature of utterances (*e.g.*, assertions, questions, directives, responses), and is the first level of analysis required by applications aiming at understanding spontaneous dialog [22,23]; (2) spoken-dialog classification, defined in [24] as the problem of assigning a task category to spoken utterances in task-oriented dialog systems; (3) dialog analysis and classification, also intended in the sense of capturing general aspects or characteristics of a dialog [25–27]. In our paper, we focus on the latter, which is the area of approaches analyzing social signals cues in the social signal processing (SSP) framework [1–3,11–15,28], in which the main concern is to understand the kind of interactions between two or more individuals. Other techniques are focused on the classification of specific aspects such as the degree of politeness, frustration, or emotional states [29]. The approaches presented in [30–33,14,34,35] have goals similar to ours: they are aimed at classifying different kinds of meeting dialogs [32] and capturing different characteristics/roles of the single participants in a conversation [30,31,33,14,34,35].

Two major issues make the dialog modeling problem hard to manage, and these concern the kind of features to be extracted and the mathematical model to be applied to process such features in order to realize the nature of the dialog.

Regarding the features, some attempts have been made to focus on one-dimensional features such as prosodic features produced in the early processing stages [36,37], and for smoothing out the dynamics of adult dialog systems [7]. Principal prosodic features are related to pitch (*e.g.*, pitch statistics, intonation patterns), energy (*e.g.*, RMS and SNR statistics), duration (*e.g.*, phonemes/words duration statistics, speech rate), and pauses (*e.g.*, statistics on their frequency and duration) [37]. Recently, prosodic features related to voice quality have also gained some attention as effective indicators of different emotional states and attitudes of the speaker [38,36], and automatic dialog analysis methods have been investigated considering emotional cues as part of prosodic information [39]. When considering the SSP field, a class of speech features specifically designed to characterize social behavior have been proposed in [34,40,15]. These are built upon low-level features, such as spectral and prosodic ones, and are used for instance to determine speaker's motivation and mental focus (*emphasis*), interest and engagement (*activity*), empathy (*mimicry*), influence of one speaker onto another (*influence*), and social control of the interaction (*dominance*).

Talkspurts (contiguous intervals of speech, with internal pauses no longer than 0.3 s [33]) and periods of silence, also known as vocal interaction features in the psycholinguistic community [41], have been widely used in conversational analysis as a mean to model the rhythmic turn-taking patterns in human–human conversations. The early approaches in [25,26] proposed a first-order, six-state Markov model that takes into account the mean duration of pauses, switching pauses (when a different speaker takes the floor), simultaneous speech, and (single speaker) vocalizations in recorded dyadic conversations. Almost in the same period, in [27], a very similar model was presented where the trained parameters were used to generate synthetic silence-speech signals which were compared to real human conversational data.

More recently, the turn-taking dynamics was refined considering the duration of the pauses as an important feature for the conversation modeling [42]. In [43], a similar approach was applied to model a conversation between a human and a synthetic agent. Talkspurts and periods of silence have also been proved to be good features for performing speaker diarization (*i.e.*, determining who spoke when) in the presence of privacy issues, *i.e.*, without considering the content of a conversation [44].

In [32], a dialog classification system is proposed, where the main task is to discriminate among three kinds of meetings. The approach is similar to ours in the sense that they study a group engaged in a discussion as a whole entity. As features, pairwise Markov probability transitions between periods of speech and silence are considered, at each time step, so that several auto-transitions can be taken into account. However, in our case, SCPs allow to select interesting time steps in which transition probabilities have to be estimated, providing less numerous but more informative features. In [33], a very similar approach is also proposed, which is applied to different tasks such as role identification, gender and seniority classification on the AMI meeting corpus.

Finally, other features derived from talkspurts and silence periods (*e.g.*, the total number of speaking turns and the total speaking length) are successfully employed for the audio modeling of dominance [45,46]. This is a further evidence of the expressiveness of the vocal interaction features for conversation analysis, usually tackled with more elaborated, multimodal, cues.

Generative models such as Markov models, and models that are built upon Markov models [25–27,42,47,2,48,32,33,35] have achieved a prominent position in the analysis and recognition of audio sequences in several domains, most notably, speech recognition [47] and natural language processing [49], since they offer a "readable" stochastic interpretation of time series. Discriminative approaches, such as support vector machines or neural architectures, cannot offer the same capability, and are exploited with less frequency. Actually, the hidden Markov model framework is well suited to capture the temporal dynamics of dialog acts [22,23]. Regarding the modeling of the conversational dynamics, early approaches exploited Markov models using various number of states [25–27,42], but they can be applied in a profitable way to dyadic conversations: this is because each Markov state captures a joint state configuration of the speakers (for instance, speakers A and B are both speaking), and a generalization to multi-agent conversation leads to an exponential growth of the state space. In [32], a set of Markov models is studied for each possible coupling of speakers, solving thus the computational issue raised above. Anyway, in this framework there is no general model that encodes a conversation, as in our case, but a set of disjoint structures. More recently, considering the durations of the turns in a conversation, semi-Markov models have also been exploited [35]. In order to extend the analysis of dyadic exchanges to multi-person discussions, both the influence model [4,3,50] and the mixed memory Markov processes [51] have been employed as efficient tools. The latter two architectures are similar to the one we adopt in our framework, providing the main advantage of decoupling complex interactions as summations of pairs of simpler ones.

With respect to the state-of-the-art, our framework is more compact and expressive at the same time than other structures; on the one hand, the generative processing is based on first-order Markov reasoning; therefore, system learning can be pursued in an economic way (fewer training data, shorter computational time), without the risk of data overfitting, which is symptomatic of more structured models (like semi-Markov architectures [35]). In addition, the system is versatile, allowing multi-agent dialog modeling that first-order Markov machinery in general cannot afford [25–27,42]. On the other side, the SCPs are expressive features that compensate the simplicity of the generative framework, helping it in modeling in a compact but expressive way the turn-taking dynamics. This leads to classification performances superior to other systems with similar generative structures [3], or exploiting different features [37,60].

3. Mathematical background

3.1. Markov models

Markov models offer a stochastic interpretation of time series in that, as well known, the next event or observation has a probabilistic dependency on the past k observations. The most trivial Markov model is a Markov chain, a simple discrete time process described by a set of N states. We denote the state variable of the system by $S_t \in \{1,\ldots,N\}$, and $P(S_t|S_{t-1},S_{t-2}\ldots,S_{t-k})$ indicates the transition probability for a model of order k. The latter is the probability of having state S_t given the previously observed sequence of states $\langle S_{t-1},S_{t-2},\ldots,S_{t-k}\rangle$ extending backward in time.

An ergodic Markov chain of order $k=1$ is formally defined as a couple $\lambda=\langle A,\pi\rangle$. A is the time-invariant transition matrix $A=\{a_{ij}\}$, where $a_{ij}\geq 0$ and $\sum_{j=1}^{N} a_{ij}=1$ represent the probability of going from state i to state j, *i.e.*, $P(S_t=j|S_{t-1}=i)$ with $i,j\in\{1,\ldots,N\}$. The initial state probability distribution $\pi=\{\pi_i\}$ represents the probability of the first state $\pi_i=P(S_1=i)$.

The parameters of a Markov chain can be easily estimated by frequency counts directly from "training" state sequences that represent meaningfully and reasonably various instances of the phenomenon to be modeled. Once the model is trained, it can act as a classifier on novel, unseen state sequences. This can be exploited by calculating the likelihood probability $P(S|\lambda)$, where $S=\langle S_1,\ldots,S_T\rangle$, with T being the length of the sequence [16]. In case several models are trained, say U models indexed by u, and we are looking for the one that best represents the observed state sequence, the maximum likelihood classification can be employed, which selects the $\tilde{u}$−th model for which

$$\tilde{u}=\operatorname*{argmax}_{u} P(S|\lambda_u) \tag{1}$$

Using a Markov model for coupled processes, where a time series has several components, is also straightforward at the conceptual level. The most obvious way of modeling coupled processes in a componential series is by defining a state space, where each state is the Cartesian product of the components (see Fig. 1a). For a C-component series, we have a *composite* state which is composed by C single-process states $\tilde{S}_t=\langle {}^1S_t,{}^2S_t,\ldots,{}^CS_t\rangle$, where the apex c indicates the c-th component process.

In this case, the state space explodes, leading to transition matrices of $N^C\times N^C$. In practice, each composition of single-process states is considered as jointly conditioned by the state configuration of the previous time step. This is the maximal amount of information that can be encoded in a composite first-order Markov process, but the exponential explosion in the size of the state space discourages such type of modeling. Moreover, a surfeit of parameters leads to overfitting, and there is often a lack of data for a large number of states, usually resulting in undersampling and numerical underflow errors.

Another, more convenient, factorization of coupled Markov process states instantiates a conditional probability of a *single*-process state with respect to *all* single-process states at the previous time step. Technically, it considers the k-th order Markov model, where the $k=C$ previous states are represented by the C states observed at time $t-1$, realizing thus the (full)

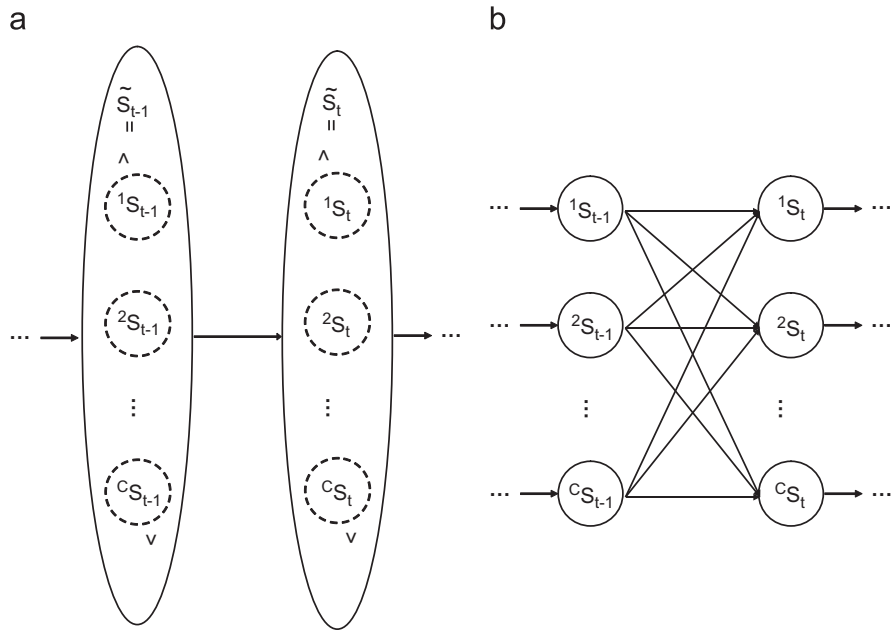

Fig. 1. Coupling of Markov models, where the arrows indicate conditional dependences: (a) Cartesian coupling of Markov models: the solid blobs represent composite states; (b) state factorization exploited by the observed influence model.

multi-process transition probability

$$P({}^{c}S_t|{}^{1}S_{t-1},\ldots,{}^{C}S_{t-1}) \tag{2}$$

leading to C different transition matrices, each of $N^C \times N$ entries. Intuitively, this factorization explains a single state as the consequence of all the component processes at the previous time step, thus modeling the effect of a "choral" conditioning.

3.2. The observed influence model

The observed influence model (OIM) has been introduced in [52] as a simplified version of the influence model [4]. OIM represents a statistical model for describing the connections between C Markov chains with a simple parametrization in terms of the "influence" each chain has on the others. The factorization of the (full) multi-process transition probability is

$$P({}^{c}S_t|{}^{1}S_{t-1},\ldots,{}^{C}S_{t-1}) = \sum_{d=1}^{C} {}^{(c,d)}\theta P({}^{c}S_t|{}^{d}S_{t-1}) \tag{3}$$

with $1 \le c,d \le C$, ${}^{(c,d)}\theta \ge 0, \sum_{d=1}^{C} {}^{(c,d)}\theta = 1$. In practice, the OIM models the full transition with a linear combination of pairwise *inter-chain* ($c \neq d$) and *intra-chain* ($c=d$) transition probabilities. The weight ${}^{(c,d)}\theta$ represents the influence that chain d exerts on chain c (a graphical representation of the state factorization is depicted in Fig. 1b).

Formally, we name an influence model as $\lambda = \{\{A^{(c,d)}\},\Theta,\pi\}$, where $A^{(c,d)}$ is the *intra*-chain matrix when $c=d$, and represents the dynamics of a single process *per se*; when $c \neq d$ we consider the *inter*-chain matrices, modeling how much a state of a chain influences the next state of the other chain. The $C \times C$ matrix Θ contains the influence weights, and π contains the (independent) initial probability distributions for all processes.

The OIM transition factorization is a good compromise between the number of parameters required ($C^2 \times N^2 + C^2$, $C^2 \times N^2$ for the transition tables parameters, and C^2 for the influence coefficients) and the expressivity of the model. In practice, the OIM is able to model each interaction between pairs of chains, but it is not able to model the joint effect of multiple chains together. In other words, $\{\theta\}$ coefficients are constant factors that tell us how much the state transitions of a given chain depend on a given neighbor.

It is important to realize the consequences of these factors being constant; intuitively, it means that how much we are influenced by a process is constant, but how we are influenced by it depends on its state.

OIM learning of the $\{\theta\}$ coefficients is performed by standard constrained gradient descent [3,16].

A classification involving the OIM has to be carried out considering carefully the order with which the observation sequences are organized. With a two-process situation in which the second process exerts a strong influence on the other, we learn a model where the weight ${}^{(1,2)}\theta$ is high. In order to recognize such situation in a classification scenario, the relative ordering of the sequences has to be preserved, *i.e.*, the first sequence has to be the one related to the process that influences the opposite one. If this cannot be ensured, a reasonable strategy for extracting the "correct" classification score would be the following: the sequences $\mathbf{S} = \{S_1,\ldots,S_C\}$ are presented to the model in all their possible orderings, indexed by o, collecting all corresponding likelihood scores $P(\mathbf{S}_o|\lambda)$; the correct likelihood score would thus be the highest one.

4. The proposed framework

For the sake of clarity, we focused here on two-person conversations, played by subjects 1 and 2, each one equipped with a microphone and a headphone. Note that our model generalize to multi-person negotiations, as shown in the experiments. The conversation originates a couple of synchronized audio signals sampled at 44 100 Hz, each one conveying the voice of a single speaker. Source separation issues were avoided by separating the players by means of a glass pane, in an adequate anechoic soundproof booth.[4] The audio signals were filtered in

[4] A different experimental environment will be presented in Section 5.

order to prune out artifacts due to unexperienced subjects, for example, due to variations in the distance from the microphone. Then, the short-term energy of the speech signals was computed on frames of 10 ms, and a speech/silence classification was performed on the energy contour by a clustering process adopting the k-means procedure [16], setting the number of clusters to 2, so as to obtain two binary arrays O_1 and O_2, of length T. A sketch of this operation is shown in Fig. 2a.

In this work we assume the two streams as originating from two cooperating binary stochastic processes. Our idea is to introduce a model which encodes the mechanism that causes one process to change or remain steady in its state, depending on its previous state and on the previous state of the other process. A simple choice could be to fit an OIM, supposing that each silence/speech sample amounts to the observation of a Markov process [53].

Looking at Fig. 2a, we can get an idea of the expected resulting transition matrices: being the silence/speech (and vice versa) switches rarer than the persistences of the signals in the same state, the resulting Markov matrices are strongly diagonal; in other words, the auto-transitions overwhelm the other transitions.

An alternative choice could be to taken into account the duration of each speech/silence segment, as an indicator of the state of each stochastic process. This brings up two issues: (1) an explosion of the space state, being present one state for each possible duration of a speech/silence period; (2) a problem in evaluating inter-chain conditional dependencies. While the first problem can be solved by employing hidden Markov models [47] that group similar durations as expression of the same (hidden) Markov state, the second issue still remains hard to tackle. As visible in Fig. 2b, it becomes difficult to evaluate the conditional dependency of a state given the other, due to problems of transition synchronization.

The proposed solution assumes that whenever a process changes its state, it causes a *global* transition that affects also the other, opposite process, injecting a novel auto-transition state (see Fig. 2c). The fragmentation caused by global transitions forces synchronization between the processes, creating $\tilde{T} < T$ different audio segments, called *steady conversational periods* (SCP), ${}^{c}I_{\tilde{t}}$, where the apex $c \in 1,2$ indexes the speaker and $\tilde{t} = 1, \ldots, \tilde{T}$ enumerates the different SCPs.

As already mentioned, the creation of the SCPs as a relevant index is motivated by several reasons, not restricted to a mere algebraic point of view. At a basic level, they encode the respiratory kinematics which determine the dynamics of self-initiated speech. They also reflect the adaptation to ongoing other-initiated speech and its dynamics [54]. These kinematics provide the basis for the coordination of prosodic and syntactic planning [55]. SCPs are meant as a first effort in modeling the real-time interplay of physiological, neuropsychological and intentional factors which determine the dynamics of speech alternation in a dialog. SCPs represent not only the alternation of speech and silence segments within a speaker, but also include turn-taking strategies, usually negotiated via audiovisual intentional cues [56].

The introduction of SCPs in our model makes it feasible to evaluate first-order intra- and inter-chain conditional probabilities (red arrows in Fig. 2c). In order to take into account the different durations of each silence and speech segment, we pooled together all the SCPs related to the speech and the "silence", respectively, so as to generate SCP histograms (see Fig. 3b).

Here, we consider the histograms as multimodal distributions that associate to each SCP value a probability of being produced in a conversation. In order to obtain a smaller state space, we decided to quantize the possible SCP values into a restrict set of durations, adopting two labels, for the short and long durations of speech, respectively. In the same way, we also quantized the durations of the SCPs related to silence. Quantization is performed by Gaussian clustering [57], which has been applied in two steps.

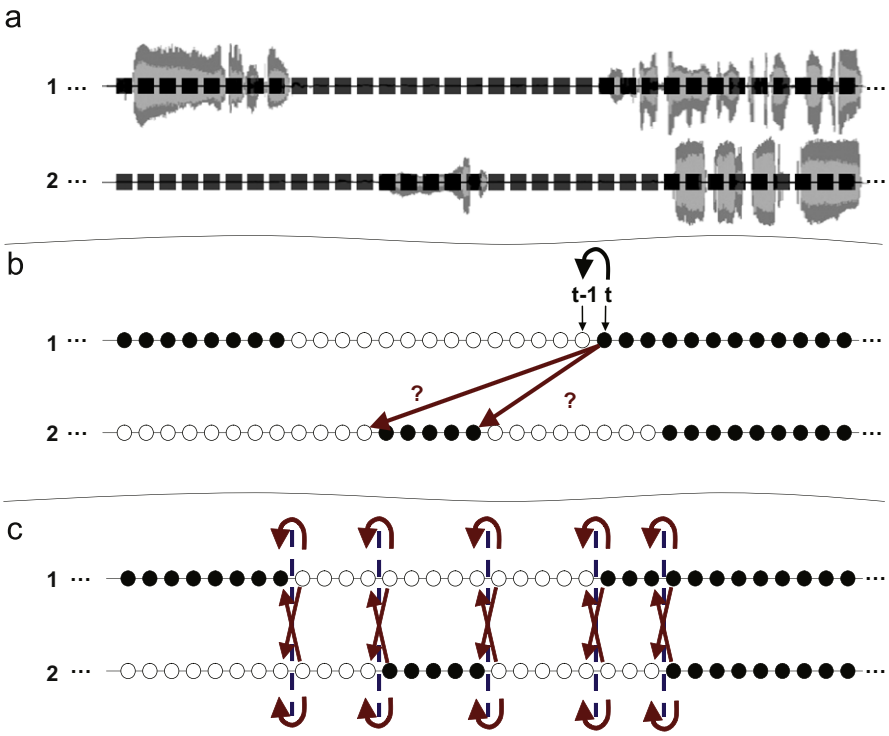

Fig. 2. Steady conversational periods creation: (a) binary conversion of the audio samples into *speech* (black dots) and *non-speech* or *silence* (white dots) values; (b) the (boundaries of the) periods of silence and speech are not synchronized, so it is not possible to evaluate a first-order statistical transition probability among periods; (c) forced synchronization due to the steady conversational periods: the synchronization permits to calculate transition probabilities *intra*- and *inter*-processes (see text). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

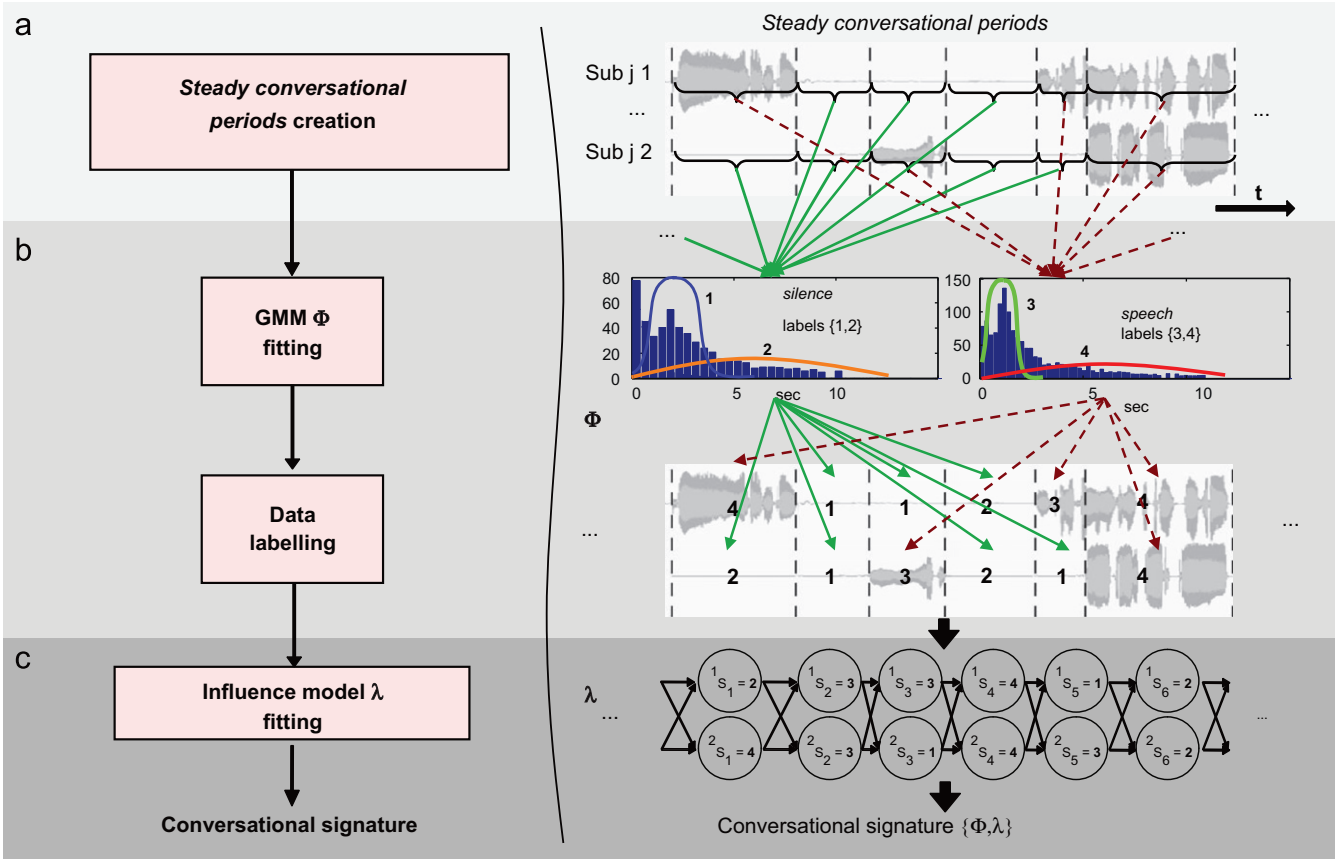

Fig. 3. Overview of the system.

In the first, we assumed that the probability of observing an SCP value ${}^{c}I_{\tilde{t}}$ follows a mixture of Gaussian (MOG) distribution, *i.e.*,

$$P({}^{c}I_{\tilde{t}}) = \sum_{r=1}^{R} w^{r}\mathcal{N}({}^{c}I_{\tilde{t}}|\mu^{r},\sigma^{r}) \tag{4}$$

where w^r, μ^r and σ^r are the mixing coefficients, the mean, and the standard deviation, respectively, of the r-th Gaussian of the mixture, and $R=2$ (short, long). We formally indicate a MOG as the set of its parameters, *i.e.*, $\Phi=\{w^r,\mu^r,\sigma^r\}_{r=1,\ldots,R}$. More specifically, we employed two GMMs, one for the SCPs related to the speech, and the other for the SCPs related to silence. The parameters of the two MOGs are estimated on training data by the expectation maximization (EM) estimation procedure [57]. Having two mixtures, we name their components univocally as $1,2,\ldots,2R$, where the first half addresses the silence SCPs, and the second half indexes the speech SCPs. The second step of the clustering imposes to assign a single Gaussian component to each SCP value. This is performed by maximum likelihood classification, *i.e.*, selecting the "nearest" (in a probabilistic sense) component of the mixture or *SCP state*, that we name ${}^{c}S_{\tilde{t}}$

$${}^{c}S_{\tilde{t}} = \operatorname*{argmax}_{r} P({}^{c}I_{\tilde{t}}|r) = \operatorname*{argmax}_{r} w^{r}\mathcal{N}({}^{c}I_{\tilde{t}}|\mu^{r},\sigma^{r}) \tag{5}$$

After this operation, each SCP state ${}^{c}S_{\tilde{t}}$ takes one label among $1,2,\ldots,2R$ (See Fig. 3b, bottom). In our previous study [58] we modeled the SCP histograms as hierarchical mixtures of Gaussians: in that way, however, we faced several problems of overfitting.

It is worth noting that the use of Gaussian clustering is motivated by the fact that we do not associate any additional discriminating information to the nature of the signal (*i.e.*, we do not perform prosodic or phonetic analysis). Therefore, context-specific clustering strategies, such as granulation [59], are not convenient here. Instead, our purpose is to build contextual information starting from low-level cues, as the SCPs are.

After the clustering, we have all the conditions that allow the modeling through the observed influence model, that is, two synchronized, discrete and inter-communicating processes. We thus fit an observed influence model $\lambda=\{\{A^{(c,d)}\},\Theta,\pi\}$ to the data.

The resulting intra-chain transition parameters indicate the conversational trend of each subject considered separately. The inter-chain transition parameters indicate *local* state dependencies among processes, while influence factors mirror the influence that a process exerts on the other. All the parameters $\{\Phi,\lambda\}$ form the statistical signature of a conversation, that will lead to an interesting analysis and classification tool.

Notice that our framework adopts a choice which extends the one proposed in [3], concerning the turn-taking modeling of dialog situations. In their work they explicitly remove temporal information regarding the persistence of a subject in a silence or speech state, while in our framework this information is carefully included in the modeling.

5. Experiments

The experimental section is subdivided in two parts.

In the first one, we will show how the statistical signature provided by the model parameters is intelligible and meaningful, explaining also subtle turn-taking aspects. In particular, we focused on a particular setting in which two kinds of dialog are taken into account, *i.e.*, dialog between adults, and between a preschool (4–6 years) child and an adult.

The study of the conversational dynamics of exchanges involving children is of great interest for several reasons. Conversations constitute complex human activities integrating executive skills and emotional resonance. Becoming a competent conversational partner

requires the acquisition in time of culture- and language-specific schemata, sustained by the development of neural networks for audiovisual comprehension of speech. It follows that a marked, biologically grounded difference is to be expected between the conversations of children and those of adults. In our experiment we assumed that time differences in the parameters of speech and silence could represent a correlate of the ability to timely integrate speech information online. Such ability is likely to be governed by the executive system, which in humans start to develop at about 6 years of age and reach full efficiency at a young adult age. Following the argument above, determining objective conversational patterns in children can be of great prospective importance for clinical work with developmental populations. Specifically, speech and communication difficulties in children are often diagnosed and assessed after treatment using subjective measures. Further, such measures usually consider only formal speech skills, and ignore the ability to interact verbally, which is instead the basis for a normal social development. Conversely, our classification scheme is based on fully objective and simple indices which nonetheless are liable to capture the core structure of conversational interaction, thereby providing a potentially optimal parameter to measure speech disorders and recovery after treatment.

In the second part, we show how our model is effective in different classification tasks, considering the adult–children dialogs, a novel dataset of dyadic conversations, and the AMI meeting corpus. In the latter case, we also show the capability of our approach to generalize to multi-person scenarios, in comparison with other approaches too.

5.1. Analysis of the model parameters: adult–adult and adult–child dialogs

27 healthy subjects (10 males, 17 females) participated in the study. They belonged to two age groups, 14 children ranging from 4 to 6 years old (average age: 5 years and 4 months), and 13 adults ranging from 22 to 40 years old (average age: 32 years old). Our dataset was composed by 27 conversational samples collected using a two-person semi-structured conversation, for which the moderator, a research-trained female psychologist who was blind to the aim of the experiment, introduced in sequence 5 predetermined topics with fixed questions in a given order (school time, hobbies, friends, food, family). Children were preliminarily made familiar with the adult moderator who was introduced to them in the presence of their parents. Written informed consent was obtained from all adult participants, and from the parents of participating children. The samples were organized into two conversational classes: (1) CH: dialog between a child and adult (14 samples), (2) AD: dialog between adults (13 samples). Each sample lasted about 10 min.

After noise removal and binarization of the conversational samples, we extracted the steady conversational periods. In Fig. 4, we show the average (*i.e.*, mediated over all the conversational

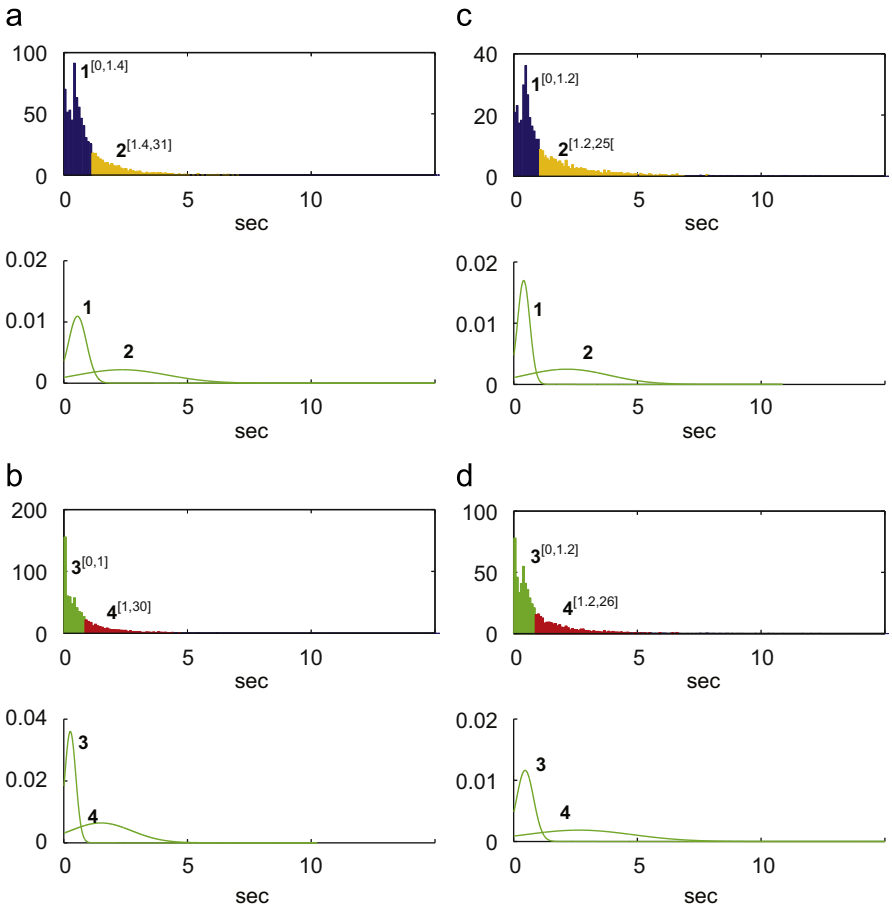

Fig. 4. Histograms of the steady conversational periods of the two dialog scenarios. (a) and (b) represent silence and speech histograms of the child SCPs for the adult-child dialog, respectively. (c) and (d) depict silence and speech histograms of the SCPs for the adult-child dialog. Underneath each of the four histograms, we also report the related Gaussian clustering, whose modes quantize in a single label the related (in a likelihood sense) SCPs. The quantization is illustrated explicitly in the histogram, where the bins related to a single mode are colored with a single tone. Near each mode, the corresponding interval boundaries that determine the quantization, expressed in sec., are reported. This gives an idea of the range of values assumed by the SCP states of the related dialog scenarios. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

samples) histograms of the SCPs forming (1) the conversations involving the children, and (2) those involving adult subjects. In general, short periods of silence and speech are shorter than 1.5 s.

The highest number of short speech SCPs is produced in the adult–child conversations (more than 150 per conversation). In general, larger SCPs (related both to silence or speech) are produced within the child dialogs. The Gaussian components fitted on the histograms highlight such profiles. As quoted in Section 4, we employed two Gaussian components for the silence SCPs and two for the speech SCPs.

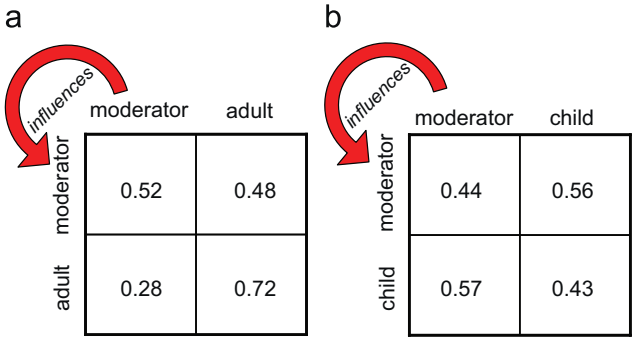

Fig. 5. Influence matrices Θ of (a) adult conversation model λ_{AD}; (b) child conversation model λ_{CH}.

After Gaussian clustering, we estimate the OIM given the labelled sequences ${}^{c}S_{\tilde{t}} \in \{1,2,3,4\}$, with the labels 1, 2 indicating short and long SCPs of silence, respectively, and the same applies for 3, 4 concerning the speech SCPs. During the training we maintained the distinction between the roles of the moderator of the conversation and the subject of the experiment (child or adult) by identifying the second sequence of the two-person dialog as the audio signal produced by the moderator.

The resulting models λ_{CH},λ_{AD} have then been employed for the classification task. The first parameters analysis, shown in Fig. 5, regards the influence matrices Θ related to the two conversational models λ_{CH},λ_{AD}.

Influence factors $\Theta = \{{}^{(c,d)}\theta\}_{c,d=\{1,2\}}$ indicate how much the subject i influences the subject j. More intuitively, a high influence between different subjects (*i.e.*, $\{{}^{(c,d)}\theta\}$ with $c \neq d$) highlights when the inter-chain probabilities regulating the choice of a state of the influenced subject given the state of the influencing subject at the previous time step, are more peaked (*i.e.*, there is a high confidence about the choice of a precise state) than the probabilities occurring in the intra-chain matrices, which exhibit uncertainty regarding the choice of a particular state of the influenced subject. One can notice from Fig. 5 that the child is more influenced by the moderator as compared to the level of influence of the moderator on the adult subject. In other words, the adult subject self-organizes his speech while the child seems to rely on the moderator to have the conversation going. In such a situation, the

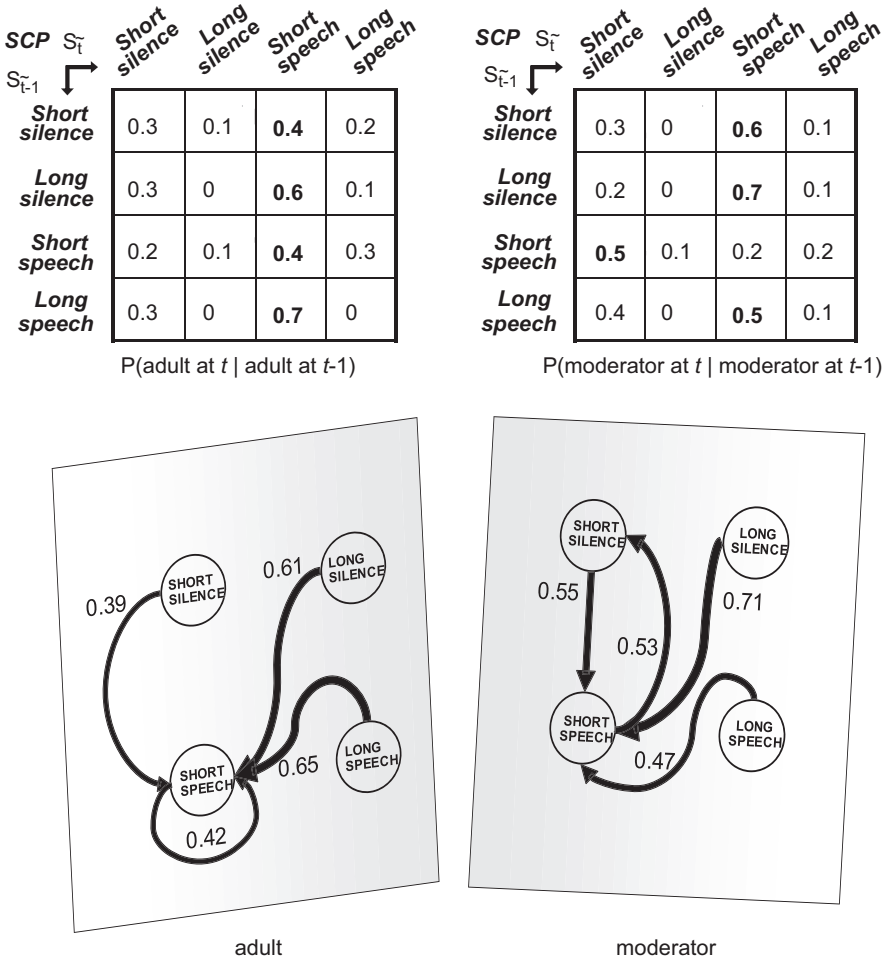

Fig. 6. Intra-chain transition matrix of the conversation between adults, and its network simplification. In the matrices, the probability values are opportunely rounded, for the sake of clarity.

two parties depend on each other and, in this perspective, the child's silence segments appear to have a communicative value.

The intra- and inter-chain transition matrices related to the adult–adult conversations and adult–child conversations are reported in Figs. 6–9. As already explained, intra-chain matrices express the first-order Markov conversational dynamics of a single subject, while the inter-chain matrices encode the probability that a particular state influences the choice of the next state of the other subject.

The figures show the values of the matrices, and portray a complementary network scheme in which circles represent states, and oriented edges conditional probabilities. From each state the most probable transition is depicted as a departing arrow, in order to allow a snapshot of the most probable paths among states that a subject may follow. The thickness of each arrow is proportional to its conditional probability. The figures portraying inter-chain matrices extend the complementary scheme by adding also the most probable inter-chain dependencies, encoded as gray arrows. For the sake of clarity, only one arrow departs from each state.

5.1.1. Dialog between adults

The intra-chain transition matrices depicted in Fig. 6 evidence a similar overall structure.

It can be inferred that, for both adults, remaining in states of long silence or long speech is rare, and this indicates the existence of a very dynamic conversational exchange. This conclusion is further supported by the fact that a short silence is likely to be followed by a short speech, and vice versa, as visible in the intra-transition table related to the moderator. Further, a long-silence state is followed by a short-speech state, and a long-speech state is followed by a short-speech state. The latter result finds an intuitive explanation after examining the inter-chain matrices, depicted in Fig. 7, and discussed in the following.

The adult–adult conversation is generally described by a high probability that all the possible states assumed by the adult are followed by a short-speech state of the moderator, who therefore drives the conversation, by stopping or encouraging speech production, and maintaining or changing topic (see Fig. 7). The inter-chain matrix that models the transition from the moderator to the adult subject highlights alternating ⟨long silence–short speech⟩ and ⟨short speech–short silence⟩ combinations. The transition from short silence to short silence represents the pauses needed to elaborate the next speech segment, for both subjects. The transition from long speech to short speech may have two different interpretations: a speaker decides to the other speaker after an excessively long speech segment, or he simply agrees with what has been just said.

5.1.2. Dialog between adults and children

The intra-chain transition matrices depicted in Fig. 8 also display interesting features. The child shows a high tendency to converge to a short silence state, while the dynamics of the moderator are more regular, displaying a high probability of moving from a state of silence to a speech state, either long or short, and vice versa.

In the inter-chain matrices (Fig. 9), the importance of the short-silence state as a peculiar aspect of the child's conversational dynamics is manifest; actually, almost all the states of the

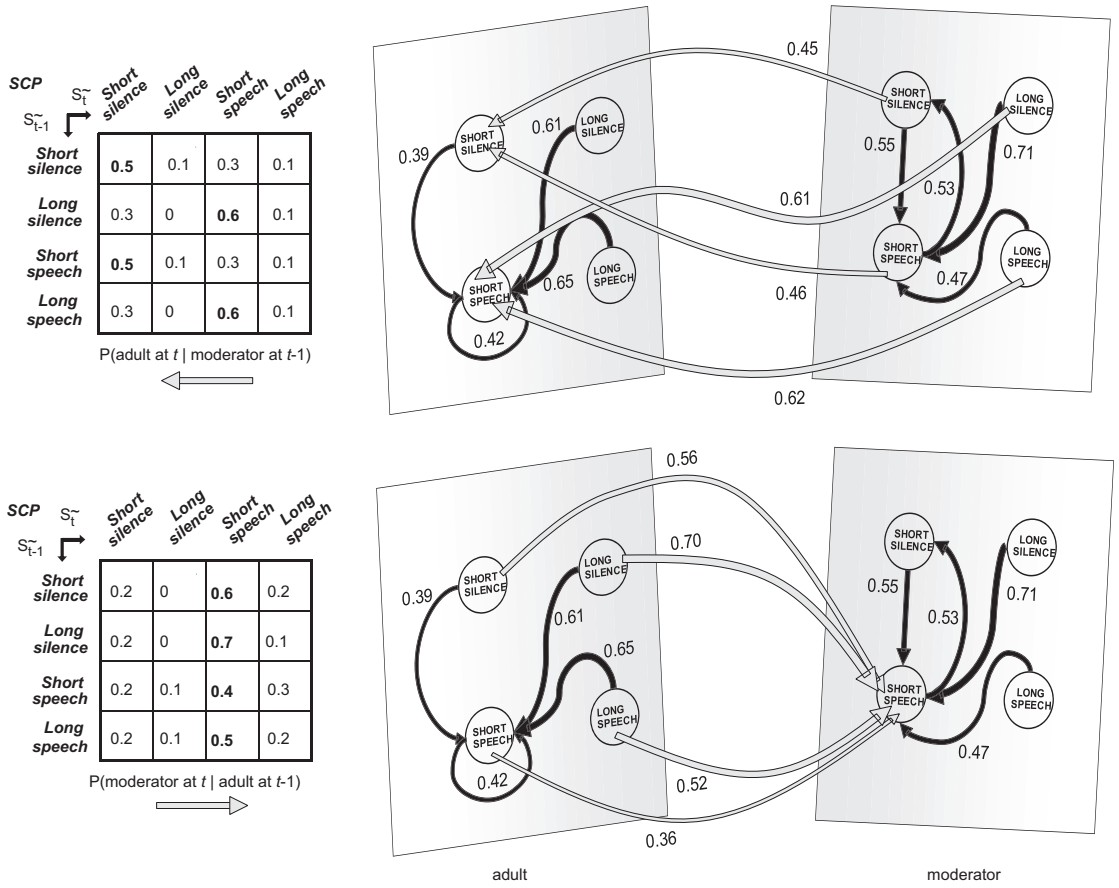

Fig. 7. Inter-chain transition matrix of the conversation between adults, and its network simplification. In the matrices, the probability values are opportunitely rounded, for the sake of clarity.

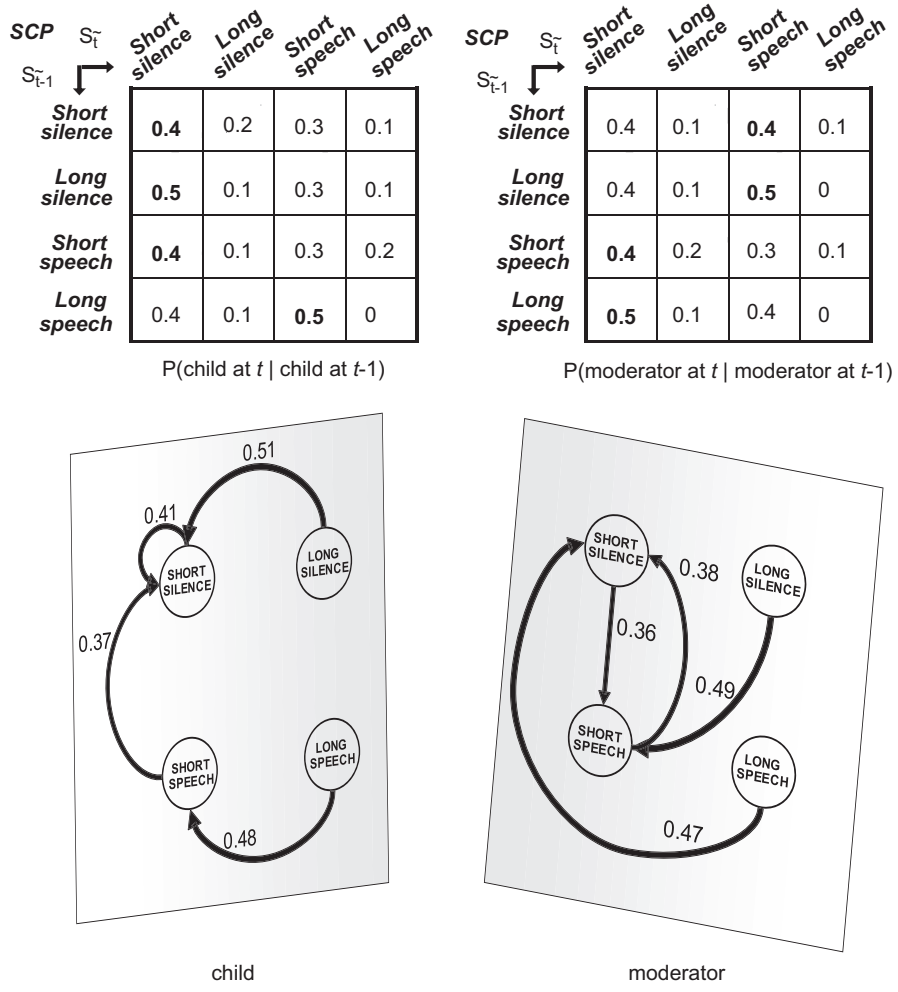

Fig. 8. Intra-chain transition matrix of the conversation between an adult and a child, and its network simplification. In the matrices, the probability values are opportunitely rounded, for the sake of clarity.

moderator are followed by a short period of silence produced by the child.

It is also worth noticing that a long speech of the moderator is followed by a short-speech segment of the child. Vice versa, the short speech and the long speech performed by the child are followed by a short period of silence of the moderator, suggesting that the moderator waits a while in order not to make the conversation too tight, thus frightening the child. A long silence of the child is followed by a moderator's short speech, which likely consists in an encouragement made by the moderator.

Two important limitations of our study must be underlined: (1) the residual rigidity of a semi-structured conversation, which may have influenced the overall conversational dynamics; (2) the emotional import of the situation, in which the child finds him/herself talking to an unknown adult about only apparently neutral topics.

5.2. Classification

We have also investigated the diverse classification capabilities of our model. This first part of this section is related to dyadic exchanges; in particular, we show how the dialog classes analyzed above (that we call *restricted* dataset) are discriminated, evaluating also classification accuracy on data portions of increasing length. Subsequently, we extend the dataset with other dyadic conversations, giving rise to different classification scenarios. All these trials are performed in a comparative way, evaluating the performances of other methods, which include baseline and advanced techniques.

The second part is related to multi-person conversations, and considers the AMI meeting corpus. In this case, we evaluated the capability of our system in profiling and discriminating the dynamics of the single participants and recognizing their role. As a comparison, we employed a technique which is very close to our framework [33].

5.2.1. Dyadic exchanges

We first present the classification results on the restricted dataset; in addition, we consider an *extended* dataset, adding to the adult conversational pool five non-structured conversations selected from a phone conversational database. The database was created collecting office conversations of our department employees, where the topics of the dialogs were focused on fixing appointments or discussions about technical information. All the conversations in this new adult conversation class are characterized by a *flat* dialog, *i.e.*, no arguing matter arose.

This was aimed at creating a more general set of adult dialogs; each sample lasted about 8 min, and the age range was the same as in the adult semi-structured conversation class. Moreover, we considered samples of phone conversations in which a dispute between adults occurred (nine samples). Such samples were created by an operator who was aware of the experimental goal, and other subjects (department employees) which were only warned about

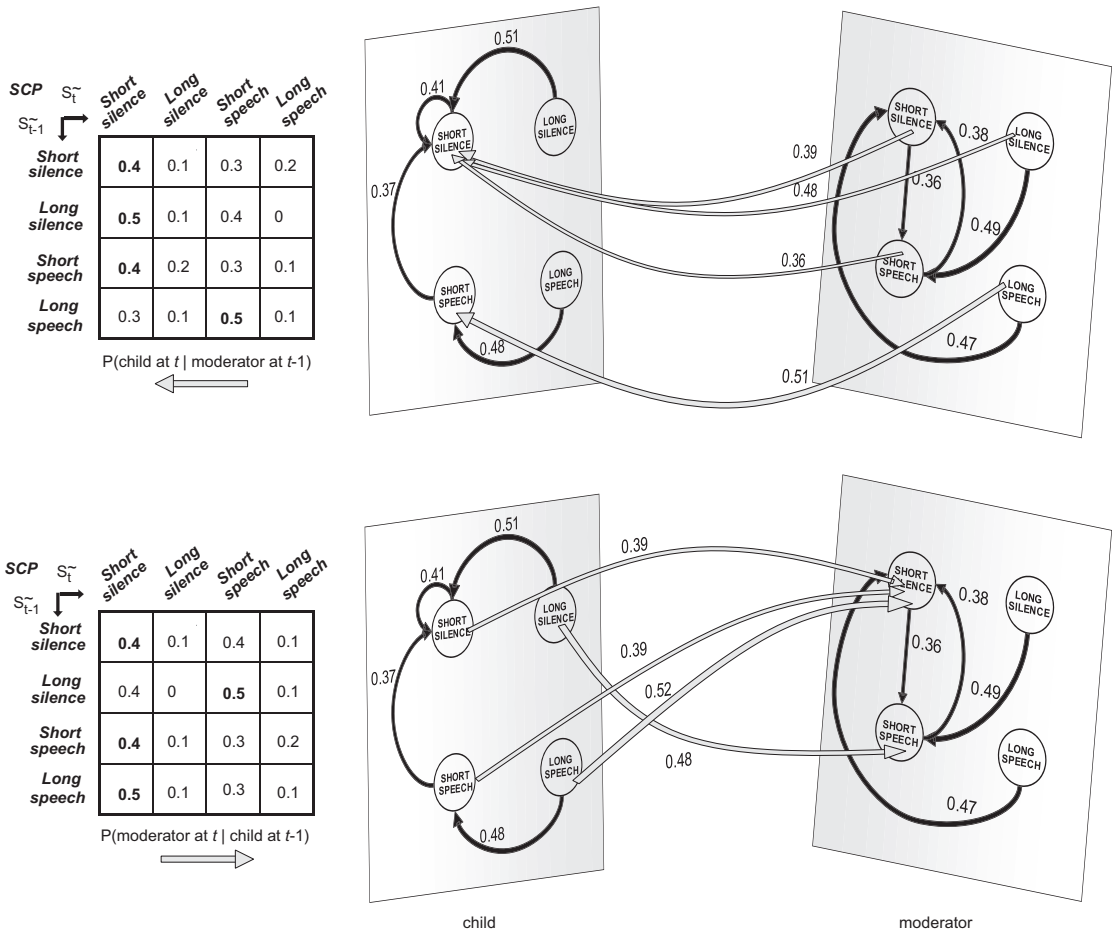

Fig. 9. Inter-chain transition matrix of the conversation between an adult and a child, and its network simplification. In the matrices, the probability values are opportunely rounded, for the sake of clarity.

the possibility that an arguing issue might arise. In this last configuration, the adult subjects ranged from 22 to 40 years (average age: 30 years) and each sample lasted about 6 min. The phone conversations were realized by recording the voice signal of each participant with a standard microphone at a sampling rate of 44 100 Hz, without relying directly on the phone signal. The signals were then synchronized. In this way, the extended dataset consists of the following three *categories* of dialogs:

1. Flat dialog between adults (18 samples).
2. Flat dialog between a child and an adult (14 samples).
3. Dispute (nine samples, only between adults).

Classification was performed in a maximum likelihood sense, as explained in Section 3, *i.e.*, learning different models, one for each category of dialog, and evaluating which one gives the highest likelihood score when fed with a test sequence. The results of the classification are obtained by cross-validation using leave-one-out [16]. The likelihood score is calculated as explained in Section 3, *i.e.*, by considering the two possible orderings of the two audio streams that compose a dialog. Classification accuracies on the restricted dataset are reported in Table 1.

Concerning model selection issues, with two Gaussian components the clusters' extrema found by the SCP clustering step were similar to those found in the previous experiments, for all the classes (short periods of speech and silence were 2 s long, while the long periods of speech and silence were more variable, depending on the dialog considered). Augmenting the number of Gaussian components to 3 (three for the silence SCPs, three for the speech SCPs), classification performances resulted similar. We also considered four Gaussian components, facing problems of overfitting, thus losing in generality and robustness of the description, other than in classification accuracy. Performances decreased severely by further augmenting the number of Gaussian components.

Table 1
Classification accuracies.

Brady	SCP	TTIM	MoG1	Joint	our approach
71%	50%	82%	75%	57%	92.5%

As a first comparative technique, we implemented the six-state Markov model of [27] (here abbreviated as *Brady*) aimed at describing the speech–silence patterns of dyadic conversational dynamics. Each person's speech is coded into 5 ms silence–speech intervals and the state transition parameters are determined by frequency count. We have two speakers, named A and B. The model for speaker A is composed by the following states:

State 1: A talks and B silent
State 2: Double talk and A is interrupted
State 3: Double talk and A is interruptor
State 4: Mutual silence and A spoke last
State 5: Mutual silence and B spoke last
State 6: B talks and A silent

The joint A–B system is represented by these six states, since each state of speaker A corresponds to a unique state for speaker

B [27]. For performing the classification in a fair fashion, we have modified the original classification scheme proposed in [27] to our setting. In [27], the task was to learn a Markov model from a training conversation O_{train}, then generate (by Monte Carlo simulation) a dialog sequence O_{gen} from it, and finally perform a comparison with three goodness-of-fit parameters between particular segments of the two sequences O_{train} and O_{gen}. The segments described 10 events, like talkspurt, mutual silence, etc.. We initially mimic as best as possible the same flowchart in our work, studying a set of models (one for each training member of a class, for all the classes) generating a sample for each training sequence, and evaluating a test sample maximizing a classification score. The classification score for a class is the best goodness-of-fit score calculated with the generated sequences of that class; as goodness-of-fit criterion we evaluate the cumulative distribution function criterion that in [27] is quoted as the most informative. We realized that this procedure has not exploited the capability of Brady's model, *i.e.*, we noted that such testing scheme was not able to exploit its generalization capability, achieving a performance to chance. Instead, we learned for each class a single Markov model from all the training sequences, and we performed the classification evaluating the simple model likelihood. In this case the performances raised, obtaining a more informative comparative test.

As second comparative strategy, we learned a pure joint state model (Table 1, *Joint*), *i.e.*, we do not apply the factorization driven by the influence factors, keeping transition matrices of $N^C \times N^C$, where $N=4$ is the number of states and $C=2$ the number of individuals involved in the conversation. The model was learned by considering the synchronization pursued by the SCPs.

As third comparative test (*SCP*), we considered only the contribution given by the SCP features for the modeling of the dialogs. In practice, we employed the normalized histograms of the silence and speech SCPs as identifiers for a particular dialog, or class of dialogs. Therefore, in order to test the class-membership of a test dialog, we simply calculated the Bhattacharyya distance between the (silence) speech SCP histogram of the test with the (silence) speech SCP histogram of the class, then multiplying the two resulting speech/silence distances and obtaining a membership score. The minimal score encodes the class-membership.

In practice, this test can be viewed as a version of the proposed system in which the turn-taking modeling by first order Markov chaining is disregarded.

As fourth comparative strategy, we learned an OIM using directly the couple of silence/speech Boolean signals as training sequence, thus originating a set of four, 2×2 transition matrices, plus a 2×2 influence matrix. After the training, the auto-transition probabilities dominate over the intra-chain matrix, reducing the significance of the resulting model, turning out in very scarce classification performances, which have been omitted. Instead, we adopted the turn-taking influence model (*TTIM*) [3], which stays in the middle between the pure OIM and our method. In practice, we selected from the couple of silence/speech signals only the four silence/speech values occurring across each global transition at time t, that is, related to ${}^1S_{\tilde{t-1}}, {}^1S_{\tilde{t}}, {}^2S_{\tilde{t-1}}, {}^2S_{\tilde{t}}$ (*i.e.*, whenever a process changes its silence/speech state: this indicates the same instants that define the SCPs). In this way, we disregard the self-similar portions of signals, learning then an OIM, so that state transitions are more informative. This strategy can be viewed as a reduced version of our model; in particular, the durations of the SCPs are omitted in this analysis.

As fifth comparative technique (*MoG1*), we considered a classifier formed by a multidimensional Gaussian trained on the values of a set of acoustic cues extracted directly from the audio streams. Such choice is consistent with the classification models reported in the literature concerning conversational speech analysis for dialog and dialog acts as classification [37,60]. The selection of the acoustic cues was made with the intention to keep the set as small as possible, yet well-matched to effectively represent our data. Since most of the acoustic cues commonly used to this aim are of a prosodic nature, we selected the pitch range measure to characterize intonation, and the "enrate" speech rate measure as a predictor of syllable articulation velocity. Both audio signals of a conversation have been employed in collecting the features to feed the classifier.

Analyzing the results in Table 1, we note that our approach reaches the highest performance. An interesting observation is that the joint factorized model provides a score which is strongly below ours. This occurred because of the complexity (number of parameters) of the model, that probably needed much more training data. This observation is validated by observing the transition matrices learned, that appear very sparse.

Considering Brady's model, the lower performance is probably due to the facts that the speakers cannot perform a state change at the same time, and, more important, that the temporal modeling of periods of silence and speech is unimodal (see [27]): this is reasonable while considering a single class of dialogs, but becomes too restrictive while observing heterogeneous exchanges.

In order to test the robustness of the model in deriving peculiar information from few data, we varied the length of the training sequences, starting from short dialog intervals, each one being extracted randomly from each sequence, and increasing incrementally their length. The trained model were then fed with the whole test sequences. The resulting accuracies are reported in Fig. 10, highlighting the fact that 50 s of conversation are enough to reach fair classification performances (around 60% of accuracy).

Considering the extended dataset, we created the following classification scenarios (where *cat.* stands for *category*):

(A) flat *vs* dispute—(*cat.*1 *vs cat.*3);
(B) flat *vs* dispute, *general*—((*cat.*1 $\cup$ *cat.*2) *vs cat.*3);
(C) with *vs* without child—(*cat.*2 *vs cat.*1);
(D) all *vs* all.

The idea here is to test the capability of the model to capture different kinds of dialog scenarios, highlighting their peculiar characteristics in terms of conversational dynamics, in order to discriminate them adequately in a classification sense. The (cross-validated) classification results are shown in Table 2.

In this case too, we compared our method with the five techniques detailed above. The only difference here concerns the *MoG* classifier, here revisited as a multidimensional Gaussian trained on the values of a larger set of acoustic cues (see Table 2, *MoG2*). Other than the previously employed pitch range measure and "enrate" speech rate, we also included the spectral flatness measure (SFM) and the drop-off of spectral energy above 1000 Hz (Do1000), two features known to be correlated to voice quality modulations observed in emotionally charged phonation [36]. This was done since our dataset included dialogs characterized by non-neutral emotional states (*i.e.*, the dispute cases). Both audio signals of a conversation have been employed in collecting the features to feed the classifier.

Our results appear promising, confirming the importance of the silence/speech alternation profile as an objective characteristic which can nonetheless provide a fine modeling of conversational behavior, in the cases of both self-organized communication and turn-taking strategies. In particular, we see that Brady's model is able to accomplish well the tasks *A* and *B*, and this probably because vocal interruptions, that intuitively characterize a dispute, are explicitly modeled as a Markov state.

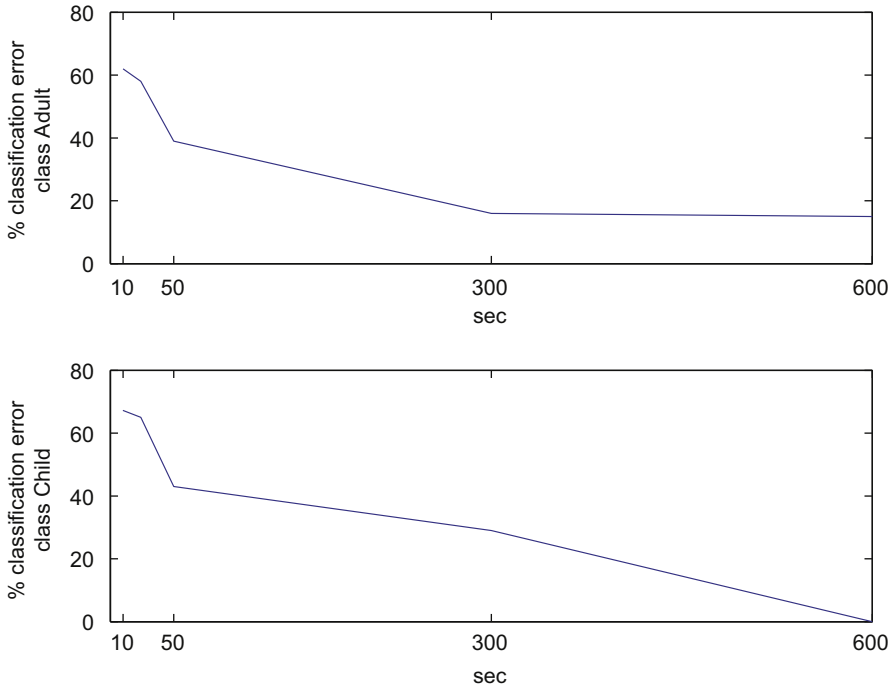

Fig. 10. Classification error when varying the length of the training sequences.

Table 2
Classification accuracies.

Scenario	Brady (%)	SCP (%)	TTIM (%)	MoG2 (%)	Joint (%)	Our approach (%)
A	80	75	100	56	53	86
B	80	79	59	54	50	86
C	50	50	64	58	52	78
D	60	50	66	64	34	80

Looking at the model parameters we have observed that transitions towards such state are more probable than in the flat negotiations.

In task *A*, our method gives lower accuracy than the TTIM model because it tends to misclassify some flat conversations. This is probably due to the fact that in some cases the timing of flat conversations uttering short speech segments, thus producing a turn-taking rhythm similar to that of dispute dialogs. This behavior is captured by our model and disregarded by TTIM. Therefore, a good future direction to investigate might be to embed features for emotion detection in conjunction with SCPs.

5.2.2. Multi-person conversations

As shown in Section 5.1, our model captures in an intelligible form the nature of the turn-taking behavior of single participant. In practice, given a speaker, its intra-chain transition matrix codifies its self-dynamics, and the inter-chain matrices encode the statistical relations with the other speakers, along with the influence coefficients. In this section, we show how this information can be exploited to segregate and discriminate precise roles assumed in a meeting. We considered a subset of the AMI Meeting Corpus [61] containing meetings recordings involving $C=4$ participants who play different specialist roles in a product design team: project manager (PM), marketing expert (ME), user interface designer (UI), or industrial designer (ID).

The corpus also provides the labels specifying the gender and the role assigned to each participant. We kept the recommended subdivision of this data into: AMITRAINSET of 98 meetings, for training the data; AMIDEVSET of 20 meetings, for defining the best features for classification; and the AMIEVALSET, 20 meetings, for performing the classification. We used the provided word alignments of these meetings as input.

In the experiments, we considered the same two tasks presented in [33] on the AMI corpus, where a model similar to ours to some extent was proposed. The tasks are (1) classifying unique roles, and (2) finding the project manager, (3) classifying the gender of speakers. The purpose of the first task is to guess the permutation of roles {PM, ME, UI, ID} of the participants, for each conversation. The second task is simpler, aimed at finding who is the project manager in the test conversations. In the third case, the goal is to guess the gender of each participant of the conversation.

For the first task, we used the training data for learning 24 models,[5] each one representing a possible permutation of roles. This was achieved simply reordering the training data accordingly. Given a test sequence, we classified the roles participants simply selecting the most probable (in a likelihood sense) of the 24 models, promoting thus one of the possible permutations. We operated directly on the AMIEVALSET, without performing feature selection in the AMIDEVSET; this is because in our case the features are the model parameters, all needed for calculating the model likelihood. As a comparison, we referred to the results shown in [33]. The similarity of their approach as compared to ours lies in a set of first-order transition probabilities over states of silence and speech, considered among speakers. Anyway, in their framework, probabilities are not exploited as a Markov model, *i.e.*, for evaluating the model likelihood, but are treated as independent features and fed into a Naive Bayes classifier.

The results are shown in Table 3. As visible, in our case we ameliorate the performances in finding two roles, decreasing that of the user interface designer. Please note, in our case we employ all the information learned during the training data, achieving

[5] In this experiment and in the following one, we set always two Gaussian components for silence/speech SCP, achieving similar clusters' extrema to those found in the previous experiments.

Table 3
Classifying unique role task: (a) the method of [33]; (b) our approach.

(a) Laskowski [33]	Hyp			
	ID	ME	PM	UI
ID	**8**	6	4	2
ME	5	**8**	4	3
PM	3	4	**12**	1
UI	4	2	0	**14**
(b) Our approach	Hyp			
	ID	ME	PM	UI
ID	**8**	4	1	7
ME	2	**12**	2	4
PM	1	2	**17**	0
UI	7	3	0	**10**

Table 4
Classification accuracies in the task "finding the manager".

Laskowski	Our approach
75%	90%

a global accuracy of 58.75%, whereas in the Laskowski approach a selection of features has to be performed to achieve 53% of accuracy. Without feature selection, Laskowski et al. reach 45% of accuracy. Therefore, we globally get higher accuracy.

For the second task, "finding the manager", we proceeded in the same modeling way as in [33]. In the training data, we collapsed each of the four-person conversations into a dyadic one, simply maintaining the silence–speech patterns of the project manager (for example, A), and collapsing the speech periods of the remaining subjects as they were a single person B. This has been done so that, in the collapsed dialog, the factitious speaker B speaks whenever at least one of the three component speakers is talking. We learn, therefore, such a "manager" model from all the collapsed training sequences.

In the classification step, from each test sample, we built four dyadic test conversations as specified above, considering each role as the project manager. Therefore, the conversation that gives the highest likelihood score with respect to the "manager" model gives the output of the classification.

Even in this case, the system presented in [33] selected some transition probabilities as useful features, while we kept all the training information for the testing. Results are shown in Table 4.

In the third case, we operated in a similar way as for the previous approach. Given a training sequence, we collapsed it in order to have four dyadic conversation. The silence–speech patterns of a person, for example A, are left unchanged, the other three are collapsed. Depending on the gender of A, for example male, the sequence will serve to learn the "male" model. In this way, we study the way a male or female faces a conversation in a group, not considering in fact the gender of the group components.

Given a test sequence, we generated four dyadic conversations as explained before, and we proceeded with the classification of the gender of the four components. In [33], only the male class was evaluated, achieving 65% accuracy. We preferred to evaluate both classes, achieving an accuracy of 81% on the male class, and 10% on the female class. As mentioned in [33], their results on this task are not statistically significant. Considering our results instead, we can say that the turn-taking patterns are not enough to gather the gender of a speaker, and other cues have to be taken into account.

Our approach and that of [33], evaluated in the three previous tasks, share some similarities, in particular they both consider transition probabilities as features. In our case however, we include also a temporal modeling which is absent in [33]. We think that this is the main contribution that allows us to get higher performances on such tasks. In other words, in modeling conversational dynamics both the succession of speech/silence patterns and their duration are fundamental parameters to consider: a very long pause has a completely different meaning with respect to a short one in the development of the vocal exchange. Using SCPs, we can embed this aspect of the conversation directly in the classifier.

6. Conclusions

In this paper, we proposed a structured generative model which, exploiting a psychologically principled low-level feature, is able to analyze conversational settings. In particular, this model is able to classify different kinds of dialog scenarios, characterized by several social situations, in an accurate manner. Our method is based on the coupling of the clustering obtained by applying the mixture of Gaussians and an observed influence model, and provides a conversational signature which is discriminant with respect to different classes of dialogs. Particularly important is the feature extraction phase, which is not based on prosodic or phonetic features typically used in classic state-of-the-art algorithms, but aims at extracting the speakers' periods of speech and silence in order to model the dynamics of the conversation employing a stochastic reasoning. More specifically, first-order Markov relations among continuous individual intervals of silence and speech are exploited; the Gaussian clustering's step permits to embed the relations in a low-dimensional state space, and the observed influence model uses such relations to encode inter-speaker interactions in a very convenient way. The feature extraction phase is easy to carry out and does not depend on contextual knowledge, (*e.g.*, the age of the speakers); instead, it furnishes cues which serve to build contextual information. Finally, we would like to stress that the nature of our framework is purely generative: this because we are interested in proposing a model which is able to provide an intuitive and readable explanation of the dialog classes, other than ensuring nice classification performances. This has also the advantage that it is not dependent on the number or the kind of classes considered.

In other words, our framework corroborates the fact that the timing of the speech/silence alternation within and between speakers is a key characteristic to consider for the interpretation of dialogs, providing a sound basis for the analysis of typical and atypical conversational behaviors.

To sum up, we proposed a behavioral blueprint of conversational skills that, for its simplicity and objectivity, may be important for tracking the variations of conversational behaviors in different settings.

Future work will be devoted to extend the experimentation to other types of dialogs, different classes of situations and to employ our framework for segmentation purposes, in which segments of a dialog were detected and analyzed in order to discover their class of membership, so as to predict the evolution of the dialog. Moreover, the two directions above will be investigated considering SCPs features in conjunction with higher level cues (*i.e.*, modeling prosody and phonetics), and in less supervised

acquisition environments. Finally, in order to increase the classification performances, our intention is to embed our generative model into a discriminative framework, *e.g.*, by adopting score space methods [62].

References

[1] A. Pentland, Social signal processing, Signal Process. Mag. IEEE 24 (4) (2007) 108–111.

[2] T. Choudhury, S. Basu, Modeling conversational dynamics as a mixed-memory Markov process, in: L.K. Saul, Y. Weiss, L. Bottou (Eds.), Advances in Neural Information Processing Systems, vol. 17, MIT Press, Cambridge, MA, 2005, pp. 281–288.

[3] S. Basu, T. Choudhury, B. Clarkson, A. Pentland, Learning human interaction with the influence model, MIT MediaLab, Technical Report 539, 2001.

[4] C. Asavathiratham, A tractable representation for the dynamics of networked Markov chain, Ph.D. Dissertation, Department of ECS, MIT, 2000.

[5] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, D. Zhang, Automatic analysis of multimodal group actions in meetings, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) .

[6] T. Jebara, Y. Ivanov, A. Rahimi, A. Pentland, Tracking conversational context for machine mediation of human discourse, in: AAAI Fall 2000 Symposium—Socially Intelligent Agents—The Human in the Loop, AAAI Press, , 2000.

[7] J. Edlund, M. Heldner, Exploring prosody in interaction control, Phonetica 62 (2005) 215–226.

[8] M. Pantic, A. Pentland, A. Nijholt, Special issue on human computing, IEEE Trans. Syst. Man Cybern. Part B 39 (1) .

[9] M. Kotti, D. Ververidis, G. Evangelopoulos, I. Panagakis, C. Kotropoulos, P. Maragos, I. Pitas, Audio-assisted movie dialogue detection, IEEE Trans. Circuits Syst. Video Technol. 18 (2008) 1618–1627.

[10] P. Dai, H. Di, L. Dong, L. Tao, G. Xu, Group interaction analysis in dynamic context, IEEE Trans Syst. Man Cybern. Part B 38 (1) (2008) 275–282.

[11] T. Choudhury, A. Pentland, Characterizing social interactions using the sociometer, in: Proceedings of NAACOS, 2004.

[12] A.S. Pentland, Socially aware computation and communication, Computer 38 (3) (2005) 33–40.

[13] J. Curhan, A. Pentland, Thin slices of negotiation: predicting outcomes from conversational dynamics within the first five minutes, J. Appl. Psychol. 92 (2007) 802–811.

[14] A. Vinciarelli, Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling, IEEE Trans. Multimedia 9 (6) (2007) 1215–1226.

[15] D. Jayagopi, H. Hung, C. Yeo, D. Gatica-Perez, Modeling dominance in group conversations using nonverbal activity cues, IEEE Trans. Audio Speech Lang. Process. 17 (3) (2009) 501–513.

[16] R. Duda, P. Hart, D. Stork, Pattern Classification, John Wiley and Sons, 2001.

[17] D. McFarland, Respiratory markers of conversational interaction, J. Speech Lang. Hear. Res. 44 (1) (2001) 128–143.

[18] S. Hurley, The shared circuits model (scm): how control, mirroring, and simulation can enable imitation, deliberation, and mindreading, Behav. Brain Sci. 31 (1) (2008) 1–22.

[19] J. Pineda, Sensorimotor cortex as a critical component of an 'extended' mirror neuron system: does it solve the development, correspondence, and control problems in mirroring?, Behav. Brain Functions 4 (1) (2008) 47.

[20] E. Nilsen, S. Graham, The relations between children's communicative perspective-taking and executive functioning, Cognitive Psychol. 58 (2) (2009) 220–249.

[21] M. Cristani, A. Pesarin, C. Drioli, A. Perina, A. Tavano, V. Murino, Auditory dialog analysis and understanding by generative modelling of interactional dynamics, in: Second IEEE Workshop on CVPR for Human Communicative Behavior Analysis, Miami, Florida, 2009.

[22] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van, E.M. Meteer, Dialogue act modeling for automatic tagging and recognition of conversational speech, Comput. Linguist. 26 (2000) 339–373.

[23] D. Surendran, G. Levow, Dialog act tagging with support vector machines and hidden markov models, in: INTERSPEECH 2006—ICSLP, 2006.

[24] C. Cortes, P. Haffner, M. Mohri, A machine learning framework for spoken-dialog classification, in: L. Rabiner, F. Juang (Eds.), Handbook on Speech Processing and Speech Communication, Part E: Speech recognition, Springer-Verlag, Heidelberg, Germany, 2008, pp. 585–595.

[25] J. Jaffe, S. Feldstein, L. Cassotta, Markovian models of dialogic time patterns, Nature 216 (1967) 93–94.

[26] J. Schwartz, J. Jaffe, Markovian prediction of sequential temporal patterns in spontaneous speech, Lang. Speech 11 (1) (1968) 27.

[27] P. Brady, A model for generating on-off speech patterns in two-way conversation, Bell Syst. Tech. J. 48 (1969) 2445–2472.

[28] A. Vinciarelli, Capturing order in social interactions, IEEE Signal Process. Mag. 26 (5) (2009) 133–137.

[29] B. Schuller, G. Rigoll, M. Lang, Hidden Markov model-based speech emotion recognition, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing 2003 (ICASSP '03), vol. 2, April 2003, pp. 1–4.

[30] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, G. Lathoud, Modeling individual and group actions in meetings with layered HMMs. IEEE Transactions on Multimedia 8(3) (2006) 509–520.

[31] M. Zancanaro, B. Lepri, F. Pianesi, Automatic detection of group functional roles in face to face interactions, in: ICMI '06: Proceedings of the Eighth International Conference on Multimodal Interfaces, ACM, New York, NY, USA, 2006, pp. 28–34.

[32] K. Laskowski, M. Ostendorf, T. Schultz, Modeling vocal interaction for text-independent classification of conversation type, in: Eighth ISCA/ACL SIGdial Workshop on Discourse and Dialogue, 2007, pp. 194–201.

[33] K. Laskowski, M. Ostendorf, T. Schultz, Modeling vocal interaction for text-independent participant characterization in multi-party conversation, in: SIGdial '08: Proceedings of the Ninth SIGdial Workshop on Discourse and Dialogue, 2008, pp. 148–155.

[34] A. Vinciarelli, M. Pantic, H. Bourlard, A. Pentland, Social signals, their function, and automatic analysis: a survey, in: IMCI '08: Proceedings of the 10th International Conference on Multimodal Interfaces, ACM, New York, NY, USA, 2008, pp. 61–68.

[35] John Grothendieck, Allen Gorin, Nash Borges, Social correlates of turn-taking behavior, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2009, pp. 4745–4748.

[36] K. Scherer, T. Johnstone, T. Bänziger, Automatic verification of emotionally stressed speakers: the problem of individual differences, in: Proceedings of the International Workshop on Speech and Computer, 1998.

[37] E. Shriberg, Can prosody aid the automatic classification of dialog acts in conversational speech?, Lang. Speech 41 (4) (1998) 439–487.

[38] C. Gobl, A. Chasaide, The role of the voice quality in communicating emotions, mood and attitude, Speech Commun. 40 (2003) 189–212.

[39] J. Liscombe, G. Riccardi, D. Hakkani-Tur, Using context to improve emotion detection in spoken dialog systems, Proceedings of the Ninth European Conference on Speech Communication and Technology EUROSPEECH'05, vol. 1, 2005, pp. 1845–1848.

[40] R. Rienks, D. Heylen, Automatic dominance detection in meetings using easily obtainable features, Revised Selected Papers of the Second Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms MLMI 2005, ser. Lecture Notes in Computer Science, vol. 3869, Springer-Verlag, 2006, pp. 76–86.

[41] J. Dabbs, R. Ruback, Dimensions of group process: amount and structure of vocal interaction, Adv. Exp. Psychol. (20) (1987) 123–169.

[42] H. Stern, S. Mahmoud, W. Kin-Kwok, A model for generating on-off patterns in conversational speech, including short silence gaps and the effects of interaction between parties, IEEE Trans. Veh. Technol. 43 (4) (1994) 1094–1100.

[43] A. Raux, M. Eskenazi, A finite-state turn-taking model for spoken dialog systems, in: NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2009, pp. 629–637.

[44] D. Wyatt, T. Choudhury, J. Bilmes, H. Kautz, A privacy-sensitive approach to modeling multi-person conversations, in: IJCAI'07: Proceedings of the 20th International Joint Conference on Artificial Intelligence, 2007, pp. 1769–1775.

[45] R. Bogdan, D. Gatica-Perez, Inferring competitive role patterns in reality TV show through nonverbal analysis, Multimedia Tools and Applications, Special Issue on Social Media (2010) 1–20.

[46] H. Hung, Y. Huang, G. Friedland, D. Gatica-Perez, Estimating dominance in multi-party meetings using speaker diarization, IEEE Transactions on Audio, Speech, and Language Processing (99) (2010) 1–1.

[47] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257–286.

[48] M. Brand, N. Oliver, S. Pentland, Coupled hidden markov models for complex action recognition, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1997.

[49] C. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge, MA, 1999.

[50] D. Zhang, D. Gatica-Perez, S. Bengio, D. Roy, Learning influence among interacting Markov chains, in: NIPS, 2005.

[51] L. Saul, M. Jordan, Mixed memory markov models: decomposing complex stochastic processes as mixtures of simpler ones, Mach. Learn. 37 (1) (1999) 75–87.

[52] C. Bishop, M. Tipping, A hierarchical latent variable model for data visualization, IEEE Trans. Pattern Anal. Mach. Intell. 20 (3) (1998) 281–293.

[53] B.C.S. Basu, T. Choudhury, A. Pentland, Towards measuring human interactions in conversational settings, in: IEEE International Workshop on Cues in Communication (CUES 2001), Hawaii, CA, 2001.

[54] D. McFarland, Respiratory markers of conversational interaction, J. Speech Lang. Hear. Res. 44 (128) (2001) 43–48.

[55] K. Hird, K. Kirsner, The relationship between prosody and discourse in spontaneous discourse, Brain Lang. 80 (2002) 536–555.

[56] D. Richardson, R. Dale, N. Kirkham, The art of conversation is coordination, Psychol. Sci. 18 (2007) 407–413.

[57] A. Dempster, N. Laird, andD. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. B 39 (1977) 1–38.

[58] A. Pesarin, M. Cristani, V. Murino, C. Drioli, A. Perina, A. Tavano, A statistical signature for automatic dialogue classification, in: Proceedings of International Conference on Pattern Recognition (ICPR 2008), 2008.

[59] W. Pedrycz, A. Gacek, Temporal granulation and its application to signal analysis, Inf. Sci. 143 (1–4) (2002) 47–71.
[60] R. Fernandez, R. Picard, Dialog act classification from prosodic features using support vector machines, in: Proceedings of Speech Prosody, 2002.
[61] J. Carletta, Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus, Lang. Resour. Eval. 41 (2) (2007) 181–190.
[62] T. Jaakkola, D. Haussler, Exploiting generative models in discriminative classifiers, in: Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II, MIT Press, Cambridge, MA, USA, 1999, pp. 487–493.

Marco Cristani is an Assistant Professor with the Department of Computer Science, University of Verona, Italy, working with the Vision, Image Processing and Sounds (VIPS) Lab and Team Leader with the Istituto Italiano di Tecnologia (IIT), Genova. His main research interests include statistical pattern recognition, generative modeling via graphical models, and non-parametric data fusion techniques, with applications to social signalling, surveillance, segmentation, and image and video retrieval.

Anna Pesarin is a Ph.D. Student with the Department of Computer Science, University of Verona, Italy, working with the Vision, Image Processing and Sounds (VIPS) Lab. His main research interests include audio and video processing, statistical pattern recognition, and generative modeling via graphical models with applications to social signalling and image and video retrieval.

Carlo Drioli is a Postdoctoral Researcher at the University of Verona, Italy, working with the Vision, Image Processing and Sounds (VIPS) Lab. His main research interests are in the field of signal processing for speech, audio and multimedia, audio and voice coding by means of physical modeling, and applications of machine learning techniques to audio and speech processing.

Alessandro Tavano is a Postdoctoral researcher at the Institute of Psychology, University of Leipzig, Germany. His current research interests focus on predictive modelling of audition, and specifically on the effects of repetition suppression and cross-modal anticipation, investigated using non-invasive brain imaging tools such as the Event Related Potentials of the EEG.

Alessandro Perina received the Ph.D. degree in Computer Science from the University of Verona with a thesis on classification with generative models. From 2006 to 2010 he has been member of the Vision, Image Processing and Sound group (VIPS) at the University of Verona. He is now a Postdoctoral researcher at Microsoft Research, Redmond working with the eScience group. His research interests are in computer vision and machine learning.

Vittorio Murino is a Full Professor with the Department of Computer Science, University of Verona, Italy, and Head of Computer Imaging facility at the Italian Institute of Technology (IIT), Genova, Italy. He is author or co-author of more than 180 papers in the fields of computer vision and pattern recognition (in particular, probabilistic techniques for image and video processing), with applications on video surveillance, biomedical image analysis, and recently, bioinformatics.