Auditory Dialog Analysis and Understanding by Generative Modelling of Interactional Dynamics

M. Cristani A. Pesarin C. Drioli A. Tavano[*] A. Perina V. Murino,[†]

Abstract

In the last few years, the interest in the analysis of human behavioral schemes has dramatically grown, in particular for the interpretation of the communication modalities called social signals. They represent well defined interaction patterns, possibly unconscious, characterizing different conversational situations and behaviors in general. In this paper, we illustrate an automatic system based on a generative structure able to analyze conversational scenarios. The generative model is composed by integrating a Gaussian mixture model and the (observed) influence model, and it is fed with a novel kind of simple low-level auditory social signals, which are termed steady conversational periods (SCPs). These 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 provide a behavioral blueprint of conversational settings without relying on segmental or continuous phonetic features. Our contribution here is to show the effectiveness of our model when applied on dialogs classification and clustering tasks, considering dialogs between adults and between children and adults, in both flat and arguing discussions, and showing excellent performances also in comparison with state-of-the-art frameworks.

1. Introduction

Social signals have been defined by psychologists as a powerful determinant of human behavior, which may have evolved as a way to establish hierarchy and group cohesion [14, 24, 16, 15]. In the recent years, there has been a growing interest in the development of the so-called conversational external mediators, i.e. dialog analysis systems that observe humans interacting with each other, capturing possible social signals and enhancing the human-human conversational exchange [15],[4]. The primary aim 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 the discussion. In this paper, we present a dialog analysis system characterizing different audio profiles among dialogic conversational situations, exploiting a novel way to encode conversational dynamics. The *key* characteristic of our approach is represented by a serial generative framework, composed by a Gaussian mixture model (GMM) [8] followed by an observed influence model [1] at the higher level. Such framework is fed by a novel type of simple, low-level auditory social signals, which are termed *steady conversational periods* (SCPs) introduced in [18]. These are built on duration of continuous slots of silence or speech, and, in addition, they take into account conversational turn-taking. This allows to easily capture and profile silence/speech dependencies in dialogs, and are motivated from a behavioral, physiological and neurophysiological level.
In practice, the system is able to capture the attitude of self-selecting for turn-taking even though the interlocutor has not yet completed his own turn. Further, it also indirectly models speech planning by characterizing the tendency to utter short sentences instead of longer propositions.

This paper contributes to the state of the art in social signal processing by showing how the proposed model collects and distills effectively the SCPs as social signals, providing a means to classify dialog instances into predetermined dialogic situations, also in comparative terms. We also analyze a clustering setting in which similar (in a psychological sense) conversational dialogs are hierarchically clustered together by using a likelihood-based similarity measure.

The rest of the paper is organized as follows. In Sect.2 a brief overview of the literature devoted to the social signal processing is presented, with emphasis on the turn-taking dynamics modelling. In Sect.3, math details are provided, in order to ease the understanding of Sect.4,

[*] A. Tavano is with the Neurorehabilitation Unit 1, Scientific Institute "E. Medea", Bosisio Parini (LC),Italy. Contacts: e-mail alessandro.tavano@bp.Lnf.it, Ph: +39 031 877250, Fax: +39 031 877499.

[†] M. Cristani, C. Drioli, A. Perina, A. Pesarin and V. Murino are with the Dipartimento di Informatica, University of Verona, Strada le Grazie 15, 37134 Verona (Italy). Contacts: M. Cristani, e-mail marco.cristani@univr.it, Tel: +39 045 8027988; V. Murino, e-mail vittorio.murino@univr.it, Tel: +39 045 8027996, Fax: +39 045 8027068. Vittorio Murino is also with the IIT, Istituto Italiano di Tecnologia Via Morego, 30 16163 Genova, Italy

where the building of our model is explained. Therefore, in Sect. 5, comparative classification and clustering results are reported, and, finally, in Sect. 6 conclusions are drawn and future perspectives are envisaged.

2. State of the art

The research on the formalization of particular social signals from a computer science perspective is an emerging field in the context of the so-called human computing or social signalling area, and a few groups, increasing in number, are devoted to investigate these issues [14, 24, 15, 17]. This research investigates the modelling of turn-taking [15, 5, 13, 6] evaluating also social signals such as activity, engagement, emphasis, and mirroring [15], detecting situations of interest, attraction [17], or dominance [20].

Two major issues make the social signaling hard to manage, *i.e.*, the kind of features to be extracted and the mathematical models to be applied in processing such features in order to extract the nature of the behavior, also capturing interacting patterns.

Regarding the features, attempts have been made at focusing on one-dimensional characteristics like the prosodic features produced in the early processing stages [21, 22] for smoothing out the dynamics of adult dialog systems [9]. Prosody is the ensemble of phonetic properties that are used in connected speech to emphasize given items or concepts, disambiguate the syntactic structure of sentences, and to express emotional states of the speaker. Prosodic features are tied to intonational, phrasing, timing and loudness variations, which have their acoustic counterpart in pitch, energy, syllable duration, and pauses [22]. Recently, prosodic features related to voice quality have also gained some attention as effective indicators of different emotional states and attitudes of the speaker [11, 21]. Automatic dialog analysis has also been investigated considering emotional cues as part of prosodic information [12].

As for the models, Markov approaches, and more complex models that build upon the Markovian paradigm [19, 6, 3] have achieved a prominent position in the analysis and recognition of audio sequences in several domains. Regarding the conversational dynamics modeling, both the influence model [1] and mixed memory Markov processes [6] have been employed as fine yet efficient tools. Such architectures are similar to the one we adopt in our framework, in the fact that they provide a mechanism for decoupling complex interactions as a weighted summation of pairs of simpler interactions.

3. Mathematical background

3.1. The Observed Influence Model

The observed influence model (OIM) has been introduced in [2] as a simplified version of the influence model. 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. We denote the state variable of a Markov chain by $S_t \in \{1, \ldots, N\}$. The factorization of the (full) multi-process transition probability of the OIM is

$$P({}^cS_t|{}^1S_{t-1},\ldots,{}^CS_{t-1}) = \sum_{d=1}^{C} {}^{(c,d)}\theta P({}^cS_t|{}^dS_{t-1}) \quad (1)$$

with $1 \leq c, d \leq C$, ${}^{(c,d)}\theta \geq 0$, $\sum_{d=1}^{C} {}^{(c,d)}\theta = 1$. 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.

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 value a_{ij}^{cd} indicates $P({}^cS_t = j|{}^dS_{t-1} = i)$. The $C \times C$ matrix Θ contains the influence weights, and π contains the (independent) initial probability distributions for all processes.

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 [1, 8].

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 second 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 correspondent likelihood scores $P(\mathbf{S}_o|\lambda)$; the correct likelihood score would thus be the highest one.

4. The proposed framework

We focused on two-person conversations, played by subjects 1 and 2, each one equipped with a microphone and a headphone. The conversation originates a couple of synchronized audio signals sampled at 44100 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. The audio signals were filtered in order to prune out noise artifacts. Then, the short-term energy of the speech signals was computed on frames of 10 msec, and a speech/silence classification was performed on the energy contour by a clustering process adopting the k-means procedure [8], 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.1a.

In this work we assume the two streams as originating from two interacting 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 a single state observation of a Markov process [20].

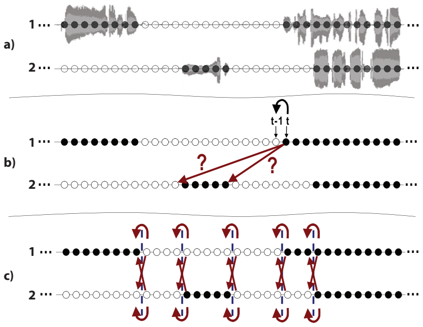


Figure 1. Steady conversational periods creation; a) binary conversion of the audio samples into *speech* (black dots) and *non speech* or *silence* (white dots) values; b) the (boundaries of the) periods of silence and speech are not synchronized, so it is not possible to evaluate a first-order statistical 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).

Looking at Fig.1a, we can get an idea of the expected resulting transition matrices: being the silence/speech (and viceversa) 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.

Another choice could be to take 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 synchrony problem in evaluating inter-chain conditional dependencies. While the first problem can be solved by employing hidden Markov models [19] that group similar durations as expression of the same (hidden) Markov state, the second issue still remains hard to tackle. As visible in Fig.1b, it becomes difficult to evaluate the conditional dependency of a state given the other, due to problems of transition synchronization.

Our solution assumes that whenever a process changes its state, it causes a *global* transition that affects also the opposite process, injecting a novel auto-transition state (see Fig.1c). 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.

The introduction of SCPs in our model makes it feasible to evaluate first-order intra- and inter-chain conditional probabilities (red arrows in Fig.1c). 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 "silence", respectively, so as to generate SCP histograms (see Fig.2b).

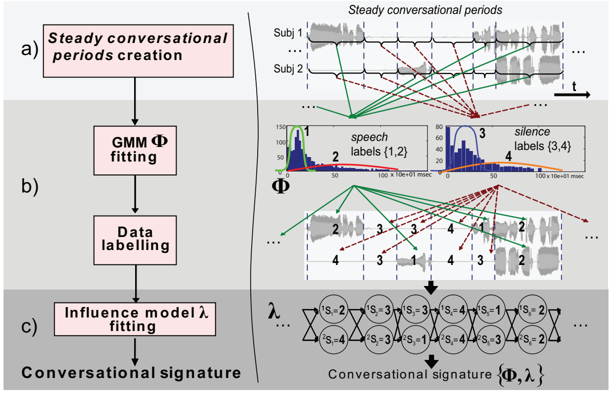


Figure 2. Overview of the system.

Here, we consider the histograms as multimodal distributions, that associate to each SCP a probability of being produced in a conversation. In order to obtain a smaller state space, we decided to quantize the possible SCP durations into a restrict set of values, 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 [7], which has been applied in two steps. 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.*,

$$P({}^{c}I_{\tilde{t}}) = \sum_{r=1}^{R} w^{r} \mathcal{N}\left({}^{c}I_{\tilde{t}} | \mu^{r}, \sigma^{r}\right) \qquad (2)$$

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.*, $\Phi = \{w^r, \mu^r, \sigma^r\}_{r=1,\dots,R}$. More specifically, we emploied 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 EM algorithm [7]. Having two mixtures, we name their components univocally as $1, 2, \dots, 2R$, where the first half addresses the speech SCPs, and the second half indexes the silence 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 ${}^cS_{\tilde{t}}$

$$ {}^cS_{\tilde{t}} \quad = \quad \arg\max_r w^r \mathcal{N}\left({}^cI_{\tilde{t}}|\mu^r, \sigma^r\right) \qquad (3)$$

After this operation, each SCP state ${}^cS_{\tilde{t}}$ takes one label among $1, 2, \dots, 2R$ (See Fig.2b, bottom).

At this point 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.

Please note that our framework adopts a choice which is orthogonal wrt what proposed in [1, 6], concerning the turn-taking modeling of dialog situations. In their work they explicitly remove the time 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 session has multiple goals. First, we would like to show how the parameters of our model are meaningful, allowing to distill intuitive behavioral patterns. Second, we will show how our model is effective in a classification task, also considering different comparative techniques. Third, we will provide results about model clustering in order to illustrate how similarity relations among dialogs can be found through their parametric modelling.
Our dialog database [1] was built considering 30 healthy subjects (12 males, 18 females). They belonged to two age groups, 14 preschool children ranging from 4 to 6 years (average age: 5 years), and 16 adults ranging from 22 to 40 years (average age: 32 years). The dataset was composed by 38 conversational samples, each sample lasting about 9 minutes.
Three categories of dialogic situations have been considered:

1. Flat dialogs between adults, formed by semi-structured (13 samples) and unstructured (5 samples) conversations. Semi-structured dialogs are driven by a moderator, a research-trained female psychologist who did not know the aim of the experiment, which introduced in sequence 5 predetermined topics with fixed questions in a given order (job/school, hobbies, friends, food, family). Unstructured conversations derived by collecting phone office conversations of our Computer Science department employees, where the topics of the dialogs were focused on fixing appointments or discussions about technical information.
2. Flat dialogs between an adult and a child (14 samples), formed by flat semi-structured conversations.
3. Dispute dialogs (6 samples), extracted by phone office conversations driven by an operator who was aware of the experimental goal, and other subjects (Computer Science department employees) which were only warned about the possibility that an arguing issue might arise.

The phone conversations have been realized by recording the voice signal of each participant with a standard microphone at a sampling rate of 44100 Hz, without relying directly on the phone signal, and the signals were then synchronized. The other face-to-face dialogs are built as described at the beginning of Sect.4. Even if the conversations have been generated in different experimental settings, the audio signal have been pre-processed/pre-filtered in order to avoid an unfair comparison among the different experimental sessions.

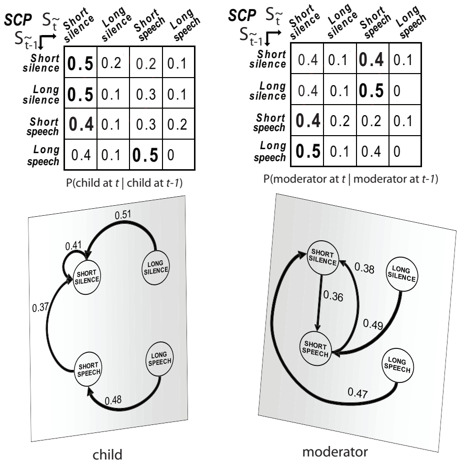
5.1. Parameters analysis: adult-child conversation

The intra and inter-chain transition matrices are reported in Fig.3, 4. As already reported, 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 conditions 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. The most probable transition is depicted as a departing arrow from each state, 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

[1] The database will be made public.

probability. The figure portraying inter-chain matrices extend the complementary scheme by also adding the most probable inter-chain dependencies, encoded as gray arrows.



P(child at t | child at t-1)

SCP $S_{\tilde{t}-1}$ \ $S_{\tilde{t}}$	Short silence	Long silence	Short speech	Long speech
Short silence	**0.5**	0.2	0.2	0.1
Long silence	**0.5**	0.1	0.3	0.1
Short speech	**0.4**	0.1	0.3	0.2
Long speech	0.4	0.1	**0.5**	0

P(moderator at t | moderator at t-1)

SCP $S_{\tilde{t}-1}$ \ $S_{\tilde{t}}$	Short silence	Long silence	Short speech	Long speech
Short silence	0.4	0.1	**0.4**	0.1
Long silence	0.4	0.1	**0.5**	0
Short speech	**0.4**	0.2	0.2	0.1
Long speech	**0.5**	0.1	0.4	0

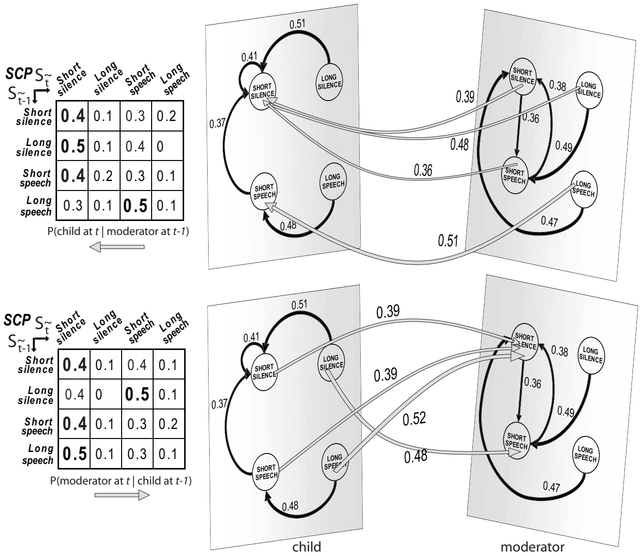
Figure 3. Intra-chain transition matrix of the conversation between an adult and a child, and its network simplification. In the matrices, the probability values are rounded, for clarity.

The intra-chain transition matrices depicted in Fig.3 display interesting features. The child shows a high tendency to converge to a short silence state, while the dynamics of the moderator is more regular, displaying a high probability of moving from a state of silence to a speech state, either long or short, and viceversa.

In the inter-chain matrices (Fig.4), 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 moderator are followed by a short period of silence of the child.

It is also worth noticing that a long speech of the moderator is followed by a short speech segment of the child. Viceversa, 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 on the part of the child is followed by a moderator's short speech, which are likely to consist in the encouragements made by the moderator.

Regarding the influence matrix, influence factors $\{^{(c,d)}\theta\}_{c,d=\{1,2\}}$ indicate how much the subject d influences the subject c. From Fig.5, one can note that the child is influenced by the moderator and viceversa, i.e., the inter-transition matrices have high importance in determining the opposite's state.



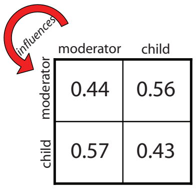
P(child at t | moderator at t-1)

SCP $S_{\tilde{t}-1}$ \ $S_{\tilde{t}}$	Short silence	Long silence	Short speech	Long speech
Short silence	**0.4**	0.1	0.3	0.2
Long silence	**0.5**	0.1	0.4	0
Short speech	**0.4**	0.2	0.3	0.1
Long speech	0.3	0.1	**0.5**	0.1

P(moderator at t | child at t-1)

SCP $S_{\tilde{t}-1}$ \ $S_{\tilde{t}}$	Short silence	Long silence	Short speech	Long speech
Short silence	**0.4**	0.1	0.4	0.1
Long silence	0.4	0	**0.5**	0.1
Short speech	**0.4**	0.1	0.3	0.2
Long speech	**0.5**	0.1	0.3	0.1

Figure 4. Inter-chain transition matrix of the conversation between an adult and a child, and its network simplification. In the matrices, the probability values are rounded, for clarity.



influences

	moderator	child
moderator	0.44	0.56
child	0.57	0.43

Figure 5. Influence matrices $\Theta = \{^{(c,d)}\theta\}$ of the child conversation model.

5.2. Classification

Classification was performed in a Maximum Likelihood sense, as explained in Sect.3, that is learning different models, one for each dialog category, 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 in a leave-one-out sense [8], i.e., by learning the model of a class with $M - 1$ samples, employing the $M-$th as test sequence and repeating this process for M times, renewing by adequate shuffling the training set. The likelihood score is calculated as explained in Sect.3, i.e., by considering the two possible ordering of the two audio streams that compose a dialog. Concerning model selection issues, augmenting the number of Gaussian components to 3 (3 for the silence SCPs and 3 for the speech SCPs), classification performances resulted similar. We also considered 4 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.

As first comparative test (see Table 1, *MG*), 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. This choice is consistent with the classification models reported in the literature concerning conversational speech analysis for dialog and dialog acts classification [22, 10]. 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 the syllable articulation velocity. 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 [21]. 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.

As second comparative technique we learn an influence model using directly the couple of silence/speech boolean signals as training sequence, originating thus a set of four, 2×2 transition matrices, plus a 2×2 influence matrix. After the training, the auto-transition probabilities dominates over the intra-chain matrix, reducing the significance of the resulting model, turning out in very scarce classification performances, which are thus omitted.

As third comparative technique, we realize an hybrid model (named here Turn-Taking Influence Model, TTIM) which stays in the middle between the pure OIM and our method. In practice, we select from the couple of silence/speech signals only the 4 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, the same instants that define the SCPs). In this way, we disregard the self-similar portions of signals, learning then an OIM. In this way, state transitions are more informative (TTIM was actually the model proposed in [6]).

The classification scenarios we took into account (where *cat.* stands for *category*) are:
(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 on Table 1.

Our results appear promising, confirming the importance of silence/speech alternation profile as an objective characteristic which can nonetheless provide a fine modeling of conversational behavior, both in the case of self-organized communication and turn-taking strategies. In the task A, our method gives lower accuracy than the TTIM model because it tends to misclassify some flat conversations. This is probably due because in some cases the timing of flat conversations is characterized by subjects which utters short sentences, 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 direction may be that of embed features for emotion detection in conjunction with SCP. In any case, the classification experiments suggest that SCPs may be promoted as effective features to be employed in modeling complex conversational behaviors.

Scenario	MG	TTIM	Our approach
A	72%	100%	86%
B	77%	62%	86%
C	58%	64%	78%
D	64%	66%	73%

Table 1. Classification accuracies.

5.3. Clustering

This section reports results about the clustering, in order to assess if it is feasible to discover natural groups of dialogs. Given the complete dataset, we perform hierarchical clustering using the complete-link scheme, employing as distance the likelihood-based similarity measure [23]:

$$D_{ij} = \frac{LL(I^i|\lambda_j) + LL(I^j|\lambda_i)}{2} \tag{4}$$

where $LL(I^i|\lambda_j)$ indicates the log-likelihood of the i-th dialog given the model $\lambda_j = \{\Phi_j, \lambda_j\}$.

In practice, we have a model for each sequence and, as in the classification task, we use a simple rule for performing clustering in order to discover the expressivity of our model. We only set the number of clusters to 3 (considering the number of categories), and let the algorithm to make the natural clusters. The resulting dendrogram is shown in Fig. 6, where in abscissas there are the category labels.

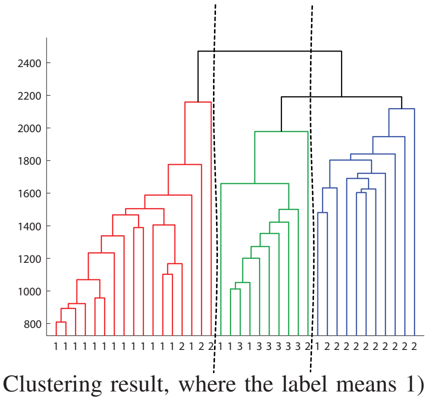


Figure 6. Clustering result, where the label means 1) adult-adult flat, 2) adult-child, 3) dispute conversations, respectively (the figure is best viewed in colors).

Observing the dendrogram, we can see that the underlying structure of the dataset is satisfactorily represented, but, obviously, there are some errors as the task is not easy. The accuracy of clustering can be quantitatively assessed by computing the number of errors: a clustering error occurs if a pattern is assigned to a cluster in which the majority of the patterns belongs to another class. In this case, we obtain a clustering accuracy of 75.63%, which is a really satisfactory result.

6. Conclusions

In this paper, we proposed a structured generative model which, exploiting a low-level yet psychologically principled social feature, is able to deal with conversational settings. In particular, this model is able to classify and cluster different kinds of dialog scenarios, characterized by different social situations (adult-adult vs. child-adult conversations, flat vs. dispute dialogs) in an accurate manner. Our method is based on the coupling of mixture of Gaussian clustering 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 first-order Markov relations.

In conclusion, we proposed a behavioral blueprint of conversational skills that, for its simplicity and objectivity, may be important for tracking the changes in time of conversational behaviors in different settings.

Future work will be devoted to extend the experimentation to other types of dialogs and different classes of situations, possibly considering more than two subjects.

References

[1] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland. Learning human interaction with the influence model. Technical Report 539, MIT MediaLab, 2001. 1, 2, 4

[2] C. Bishop and M. Tipping. A hierarchical latent variable model for data visualization. *IEEE Trans. PAMI*, 20(3):281–293, 1998. 2

[3] M. Brand, N. Oliver, and S. Pentland. Coupled hidden markov models for complex action recognition. In *Proc. of CVPR*, 1997. 2

[4] R. Brown. *Group Polarization in Social Psychology*. The MIT Press, New York, 1986. 1

[5] L. Chen, M. Harper, A. Franklin, T. R. Rose, I. Kimbara, Z. Huang, and F. Quek. A multimodal analysis of floor control in meetings. In *Proc. MLMI*, pages 263–267, 2006. 2

[6] T. Choudhury and S. Basu. Modeling conversational dynamics as a mixed memory markov process. In *Proc. NIPS*, 2004. 2, 4, 6

[7] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39:1–38, 1977. 3, 4

[8] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, 2001. 1, 2, 3, 5

[9] J. Edlund and M. Heldner. Exploring prosody in interaction control. *Phonetica*, 62:215–226, 2005. 2

[10] R. Fernandez and R. Picard. Dialog act classification from prosodic features using support vector machines. In *Proc. of Speech Prosody*, 2002. 6

[11] C. Gobl and A. Chasaide. The role of the voice quality in communicating emotions, mood and attitude. *Speech Communication*, 40:189–212, 2003. 2

[12] J. Liscombe, G. Riccardi, and D. Hakkani-Tur. Using context to improve emotion detection in spoken dialog systems. In *Proc. of EUROSPEECH*, volume 1, pages 1845–1848, 2005. 2

[13] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *IEEE Trans. PAMI*, 27(3), 2005. 2

[14] M. Pantic, A. Pentland, and A. Nijholt. Special issue on human computing. *IEEE Trans. SMC*, 39(1), 2009. 1, 2

[15] A. Pentland. Social dynamics: Signals and behavior. In *Proc. of Developmental Learn*, pages 263–267, 2004. 1, 2

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

[17] A. Pentland and A. Madan. Perception of social interest. In *In Proc. ICCV-PHI*, 2005. 2

[18] A. Pesarin, M. Cristani, V. Murino, C. Drioli, A. Perina, and A. Tavano. A statistical signature for automatic dialogue classification. In *Proc.ICPR*, 2008. 1

[19] L. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. of IEEE*, 77(2):257–286, 1989. 2, 3

[20] B. C. S. Basu, T. Choudhury and A. Pentland. Towards measuring human interactions in conversational settings. In *Proc. CUES*, Hawaii, CA, 2001. 2, 3

[21] K. Scherer, T. Johnstone, and T. Bänziger. Automatic verification of emotionally stressed speakers: The problem of individual differences. In *Proc. of SPECOM*, 1998. 2, 6

[22] E. Shriberg. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41(4):439–487, 1998. 2, 6

[23] P. Smyth. Clustering sequences with hidden Markov models. In *Proc. NIPS*. 6

[24] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland. Social signals, their function, and automatic analysis: a survey. In *IMCI*, pages 61–68, 2008. 1, 2