Generative Modelling of Dyadic Conversations: Characterization of Pragmatic Skills During Development Age

Anna Pesarin[1], Monja Tait[5], Alessandro Vinciarelli[2,3], Cristina Segalin[1], Giovanni Bilancia[5], and Marco Cristani[1,4]

[1] *University of Verona*, Italy
[2] *University of Glasgow*, UK
[3] *Idiap Research Institute, Switzerland*
[4] *Istituto Italiano di Tecnologia (IIT)*, Genova, Italy
[5] *Accademia di Neuropsicologia dello Sviluppo (A.N.Svi.)*, Parma, Italy

Abstract. This work investigates the effect of children age on pragmatic skills, i.e. on the way children participate in conversations, in particular when it comes to turn-management (who talks when and how much) and use of silences and pauses. The proposed approach combines the extraction of "Steady Conversational Periods" - time intervals during which the structure of a conversation is stable - with Observed Influence Models, Generative Score Spaces and feature selection strategies. The experiments involve 76 children split into two age groups: "pre-School" (3-4 years) and "School" (6-8 years). The statistical approach proposed in this work predicts the group each child belongs to with precision up to 85%. Furthermore, it identifies the pragmatic skills that better account for the difference between the two groups.

Keywords: Turn-Management, Conversation Analysis, Pragmatics, Social Signal Processing

1 Introduction

Pragmatics investigates "*how speakers organize what they want to say in accordance to who they're talking to, where, when and under what circumstances*" [18]. Hence, the development of pragmatic skills is a crucial step towards effective interaction with others for both humans [17] and artificial agents [3]. This work investigates pragmatic skills of children in developmental age and, in particular, it shows that statistical models of turn-management (who talks when and how much) and silence - two of the most important aspects of pragmatics - predict with satisfactory performance the age group of developing children. In other words, the work shows that age influences children pragmatics to an extent sufficient to be automatically detected with machine intelligence approaches.

The proposed approach extracts Steady Conversational Periods (SCP) [4] from conversation recordings and feeds them to Observed Influence Models (OIM) [15]. Then, it applies Generative Score Spaces (GSS) [13] and feature

selection strategies to distinguish between models trained over conversations involving children of different age groups. The experiments were performed over a corpus of 38 conversations involving two children each (76 subjects in total). Half of the conversations include children in *pre-School* (pS) age, while the other half include children in *School* (S) age. The children of the pS group are 3-4 years old, while the others are 6-8 years old.

The results show that children can be automatically assigned to the correct age group with precision up to 85%. Furthermore, the use of GSS and feature selection shows that the pragmatic aspects that better discriminate between the two age groups are (i) the probability of observing a long silence after a long period of sustained conversation, (ii) the probability of observing short periods of sustained conversation after long silences, and (iii) the probability of observing a long silence after a short period of sustained conversation. Overall, the probabilities above suggest that S children manage to sustain conversation for longer periods and more frequently than pS children.

The rest of the paper is organized as follows: Section 2 provides a brief overview of related work, Section 3 illustrates the proposed methodology, Section 4 reports on experiments and results, and Section 5 draws some conclusions.

2 Related Work

Both development and computing literature propose a large number of works where pragmatics related measurements (e.g., total speaking time, statistics of turn length, prosody, voice quality, etc.) are shown to be the evidence of social and psychological phenomena.

From a development point of view, most of the literature focused on the interaction between gestures and first words of the child, with particular attention to the phylo-ontogenetic origin of human language and its hypothetical link with the premotor system [7]. Several researchers examined the development of skills like decoding and production of pragmatic discourse parts like, e.g., intonation and verbal prosody [8, 12]. Recently, prosodic features related to voice quality have also gained some attention as effective indicators of different emotional states and attitudes of the speaker. A branch of research in fact, focuses on the evolution of conversational qualities in age of development, studying temporal features of the speech such as turns, duration, overlapping, and communication effectiveness [2].

Measurable evidences of pragmatics were extensively investigated in the computing community as well (see [16] for an extensive survey). Examples include the work in [10], where a dialogue classification system discriminates three kinds of meetings using probability transitions between periods of speech and silence, the experiments in [9], where features based on talkspurts and silence periods (e.g., the total number of speaking turns and the total speaking length) model dominance, the approach of [11], where intonation is used to detect development problems in the early childhood, and the work in [14], where prosody analysis allows the identification of language impaired children.

3 The Approach

In line with [4], the first step of the approach is the extraction of *Steady Conversation Periods* (SCP), turn management features extracted directly from audio signals: at every moment, every conversation participant i is in a state $k_i \in [0, 1]$, where 0 accounts for the participant being silent and 1 accounts for the participant speaking ($i = 1, \ldots, C$, where C is the total number of conversation participants). A SCP is the time interval between two consecutive state changes (not necessarily of the same participants). Hence, there is a sequence of SCPs for each participant i: $\{(d(n), k_i(n)\}$, where $d(n)$ is the duration of the SCP and $k_i(n)$ is the state of speaker i in SCP n. Length of the sequence and duration $d(n)$ of every sequence element are the same for all participants because the SCP changes whenever any of the participants changes state.

Overall, the extraction of the SCPs corresponds to a segmentation of the conversation into intervals during which the configuration (who talks and who is silent) is stable. In order to take into account different durations while keeping the number of states in the Observed Influence Model finite (see below), the durations $d(n)$ are grouped into two classes (*short* and *long*) by an unsupervised Gaussian clustering performed over a training dataset.

3.1 The Observed Influence Model

The Observed Influence Model (OIM) [15] is a generative model for interacting Markov chains. For a chain i ($i = 1, \ldots, C$, where C is the total number of chains), the transition probability between two consecutive states $S_i(t-1)$ and $S_i(t)$ is:

$$P(S_i(t)|S_1(t-1), \ldots, S_C(t-1)) = \sum_{j=1}^{C} {}^{(i,j)}\theta P(S_i(t)|S_j(t-1)) \tag{1}$$

where $1 \leq i, j \leq C$, ${}^{(i,j)}\theta \geq 0$, $\sum_{j=1}^{C} {}^{(i,j)}\theta = 1$, and $P(S_i(t)|S_j(t-1))$ is the probability of chain i moving to state $S_i(t)$ at step t when chain j is in state $S_j(t-1)$ at step $t-1$. An OIM can be defined as $\lambda =< A^{(i,j)}, \pi, \theta >$ ($1 \leq i, j \leq C$) where $A^{(i,j)}$ is the matrix such that $A_{kl}^{(i,j)} = P(S_i(t) = l|S_j(t-1) = k)$, π is a $C \times L$ (L is the total number of states) matrix such that $\pi_{ik} = P(S_i(1) = k)$ and θ is a $C \times C$ weights matrix where $\theta_{ij} = {}^{(i,j)}\theta$. In our case, we have dialogic conversations, i.e., $C = 2$; we have also $L = 4$ states since we have two classes (short, long) for each kind of SCP (silence, speech).

3.2 Generative Score Space

Generative Score Spaces (GSS) allow one to discriminate between generative models trained over samples belonging to different classes [13]. Their ultimate goal is to combine the advantages of both generative and discriminative approaches. In particular, the explanatory power of the parameters for the former

and the higher classification accuracy for the latter. Given a sequence of observations $\{O_t\}$, and a family of generative models $\{P(O_t|\lambda)\}$ the GSS maps the observations into a features vector ${\psi_F}^f$ of a fixed dimension for each data sample.

$$ {\psi_F}^f = F\left(f\left(\{P\left(O|\lambda\right)\}\right)\right), \tag{2} $$

where f is a function induced by generative models and F is some operator applied to it. In our case, where $C = 2$, $\{O_t\}$ is a sequence of SCPs that identifies a conversation, f is the function that estimates the transition probability matrices $A^{(i,j)}$ learned on $\{O_t\}$ [6]and F is the following operator:

$$ F\left(A_{kl}^{(i,j)}\right) = \frac{1}{2}\left(A_{kl}^{(i,j)} + A_{kl}^{(j,i)}\right) \text{ if } i \neq j; \qquad F\left(A_{kl}^{(i,i)}\right) = \frac{1}{2}\left(A_{kl}^{(1,1)} + A_{kl}^{(2,2)}\right) \tag{3} $$

It basically considers inter and intra probability values, averaging over the different speakers, reaching thus invariance with respect to the speakers order. At the end, avoiding repeated values, the feature vector ${\psi_F}^f$ has size $2L^2$.

4 Experiments

The goal of the experiments is to investigate the effect of age on pragmatic skills for children between 4 and 8 years old. The analysis includes two main steps, the first is the quantitative analysis of silence and speech, the second is a psychological interpretation of the OIM parameters after training over *pre-School* or *School* children (see below).

4.1 The Data

The corpus used for the experiments includes 38 dyadic conversations between Italian children of the same age (76 subjects in total). The corpus is split into two parts, 19 conversations involve 3-4 years old children, named *pre-School* (pS) hereafter, for a total of 38 subjects. The other 19 conversations include 6-8 years old children, named *School* (S) hereafter, for a total of another 38 subjects. The experimental setting corresponds to a *controlled observation* (see Figure 1), the children sit close to one another and fill an album, in a situation not particularly different from their everyday experience. The average duration of the conversations is 15 minutes and 31 seconds for pS children and 15 minutes and 21 seconds for S children. The conversations have been recorded with an unobtrusive Samsung Digital Camera *34*×.

Data was manually processed independently by two different annotators, in order to perform error-free source separation; as silence periods we considered segments that don't contain sounds; sounds like cough, sneezing, ambient noise. As speech, we considered all other segments that contain verbal sounds. Silences shorter than 600 ms have been considered part of a *speech* segment.

[6] We found that considering the coefficients ${}^{(i,j)}\theta$ does not help in the classification.

Fig. 1. Experimental setting. The children sit close to one another and fill an album.

Class	Silence SCP	Speech SCP
pS	74%	26%
S	72%	28%

Table 1. Amount of silence and speech SCPs for each class.

Class	Short Silence	Long Silence	Short Speech	Long Speech
pS	1.48	21.89	1.41	3.84
S	1.32	17.08	1.21	4.42

Table 2. Mean values (sec.) for short and long SCPs.

4.2 Quantitative Analysis of SCPs

The overall amount of silence and speech for pS and S conversations is reported in Table 1 and shows no significant differences between the two types of conversation. However, differences emerge when speech and silence SCPs are split into *short* and *long* classes using a Gaussian clustering (see Section 3) [5]. In particular, Table 2 shows that the mean of long silence durations is significantly higher for pS children. In other words, pS children tend to interrupt their conversations for longer periods, on average.

4.3 Classification and parameters analysis

In order to confirm the finding above, 38 OIMs were trained over the corpus, one over each conversation. The states correspond to *short* and *long* silence and speech SCPs (four states in total). The resulting OIM parameters are mapped into a score space as described in Section 3.2 so the features extracted from each conversation are the transition probabilities between OIM states (32 features in total).

Figure 2 shows the mean values of the transition probabilities for the two types of conversation. The 5 features $F = \{f_{30}, f_{14}, f_{13}, f_{27}, f_{29}\}$ were selected as those with the highest difference between pS and S children. After this manual selection, an exhaustive feature selection procedure was applied (based on all the possible combinations of features evaluated with the K-nearest neighbor classifier [6], where K was chosen with a model selection procedure using the

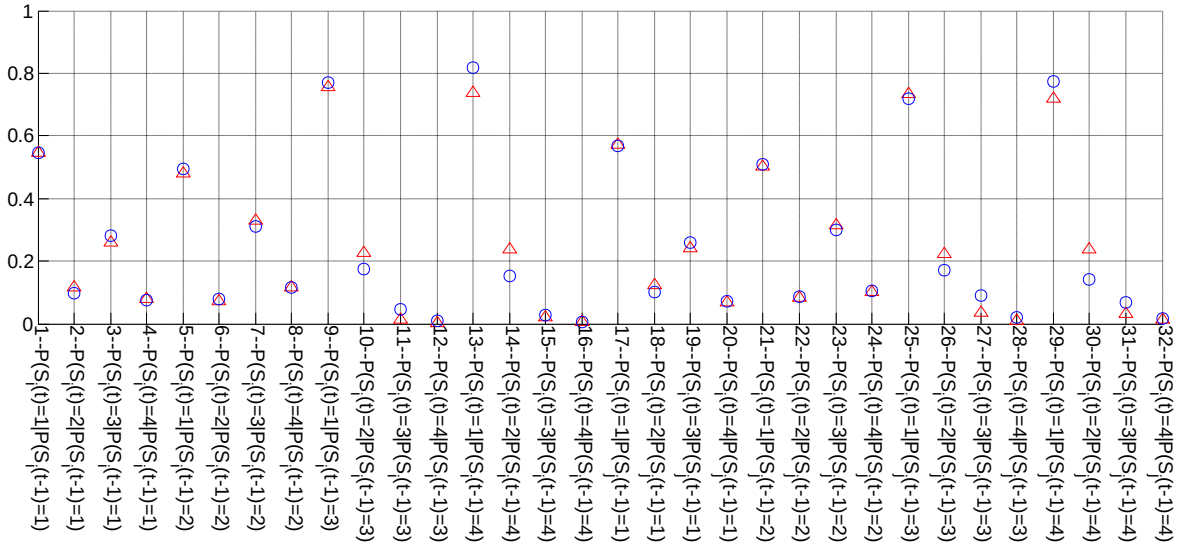

Fig. 2. Mean values of features for both classes. Triangles for pS and circles for S. Each feature has an identification number followed by its meaning.$P\left(S_i\left(t\right)=A|S_j\left(t-1\right)=B\right)$ indicate the probability of speaker S_i to be in state A after that speaker S_j was in state B at time $t-1$. The meaning of the states is: 1 = short silence, 2 = long silence, 3 = short speech, 4 = long speech.

Performances	pS class	S class
Precision	74%	87%
Recall	89%	68%

Table 3. Classification performance of the proposed approach.

PRTools toolbox [1]), that led to the final feature set used for the classification experiments: $F_b = \{f_{30}, f_{27}\}$. [7]

The K-Nearest Neighbors classifier was applied using the F_b feature set as plotted in Figure 3. The classification performances, reported in Table 3, are obtained by applying a leave-one-out approach, inserting the test sequence in the training dataset, and exploiting another sequence of the pool as test.

[7] It is worth noting that each feature of F_b, taken independently, gave rise to two sets of feature values (one for each class), which were statistically independent considering the student's t-test. In particular, the null hypothesis was rejected with a significance level of 3%.

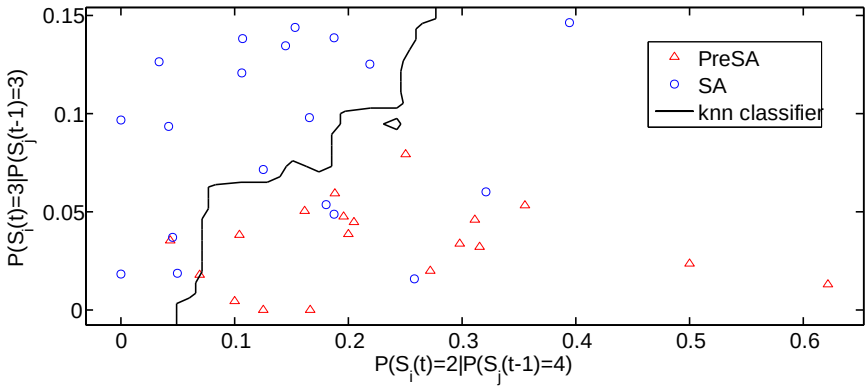

Fig. 3. The pool of conversations of the two classes as 2D points: the coordinates are the features selected by exhaustive feature selection on the set F, i.e., $f_{30} = P\left(S_i(t) = 2|P(S_j(t-1) = 4\right)$ (x-axis), $f_{27} = P\left(S_i(t) = 3|P(S_j(t-1) = 3\right)$ (y-axis).

The classification effectiveness suggests that the selected features actually characterize the two classes. The use of OIM and GSS allows one to interpret the OIM parameters under a psychological point of view. Feature f_{30} is related to the probability of transition between long speech intervals of one speaker and long silences of the other, and is higher for pS subjects.

The other feature used for the classification is f_{27}, which is an inter-speaker probability that accounts for the transition between two short speech states; its values confirm that the S conversational rhythm is higher than pS subjects. Indeed, this transition can occur when we are in presence of overlapping speech or (less frequently) an alternation of speech periods without pauses inside.

Overall, the results showed that S subjects seem to keep a higher conversational rhythm compared to pS subjects.

5 Conclusion

This paper offers a novel study of how effectively turn taking markers can discriminate the age of children. The use of Steady Conversational Periods, fed into hybrid classifiers, allowed to finely separate classes of pre-scholar and scholar conversations, explaining actually how the two classes are different: scholar children tend to have longer and more frequent periods of sustained conversation. This study promotes many future developments: considering children of different nationalities may generalize the results obtained; more importantly, this approach may lead to the definition of a clinical semeiotics able to individuate automatically pragmatic language impairments, such those that characterize autism.

References

1. Prtools version 4.1: A matlab toolbox for pattern recognition. Internet (2004), http://www.prtools.org
2. Bishop, D.V., Adams, C.: Conversational characteristics of children with semantic-pragmatic disorder. ii: What features lead to a judgement of inappropriacy? The British journal of disorders of communication 24(3), 241–263 (1989)
3. Cassell, J.: Embodied conversational interface agents. Communications of the ACM 43(4), 70–78 (2000)
4. Cristani, M., Pesarin, A., Drioli, C., Tavano, A., Perina, A., Murino, V.: Generative modeling and classification of dialogs by a low-level turn-taking feature. Pattern Recogn. 44(8) (2011)
5. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. B 39, 1–38 (1977)
6. Duda, R., Hart, P., Stork, D.: Pattern Classification. John Wiley and Sons (2001)
7. Fogassi, L., Ferrari, P.F.: Mirror neurons and the evolution of embodied language. Current Directions in Psychological Science 16(3), 136–141 (2007)
8. Friend, M.: Developmental changes in sensitivity to vocal paralanguage. Developmental Science 3(2), 148–162 (2000)
9. Hung, H., Huang, Y., Friedl, G., Gatica-perez, D.: Estimating the dominant person in multi-party conversations using speaker diarization strategies. In: ICASSP (2008)
10. Laskowski, K.: Modeling vocal interaction for text-independent classification of conversation type. In: Proc. SIGdial. pp. 194–201 (2007)
11. Mahdhaoui, A., Chetouani, M., Cassel, R., Saint-Georges, C., Parlato, E., Laznik, M., Apicella, F., Muratori, F., Maestro, S., Cohen, D.: Computerized home video detection for motherese may help to study impaired interaction between infants who become autistic and their parents. International Journal of Methods in Psychiatric Research 20, e6–e18 (2011)
12. Morton, J.B., Trehub, S.E.: Children's understanding of emotion in speech. Child Development 72(3), 834–843 (2001)
13. Perina, A., Cristani, M., Castellani, U., Murino, V., Jojic, N.: Free energy score space. In: Advances in Neural Information Processing Systems 22, pp. 1428–1436 (2009)
14. Ringeval, F., Demouy, J., Szaszák, G., Chetouani, M., Robel, L., Xavier, J., Cohen, D., Plaza, M.: Automatic intonation recognition for the prosodic assessment of language impaired children. IEEE Transactions on Audio, Speech and Language Processing 19(5), 1328–1342 (2011)
15. S. Basu, T. Choudhury, B.C., Pentland, A.: Towards measuring human interactions in conversational settings. In: IEEE Int'l Workshop on Cues in Communication (CUES 2001). Hawaii, CA (2001)
16. Vinciarelli, A., Pantic, M., Bourlard, H., Pentland, A.: Social signals, their function, and automatic analysis: a survey. In: IMCI '08: Proceedings of the 10th international conference on Multimodal interfaces (2008)
17. Wharton, T.: The Pragmatics of Non-Verbal Communication. Cambridge University Press (2009)
18. Yule, G.: Pragmatics. Oxford University Press (1996)