Exploring Children's Verbal and Acoustic Synchrony: Towards Promoting Engagement in Speech-Controlled Robot-Companion Games

Theodora Chaspari [1,2] Samer Al Moubayed [1] Jill Fain Lehman [1]
[1] Disney Research, Pittsburgh, PA, USA
[2] University of Southern California, Los Angeles, CA, USA
{theodora.chaspari, samer.-nd, jill.lehman}@disneyresearch.com

ABSTRACT

Children's interpersonal synchrony has been related to various benefits in social, mental and emotional development. We explore verbal and acoustic synchrony patterns between pairs of children playing a speech-controlled video game. Verbal features include word timing and duration patterns, while acoustic cues contain prosodic information. Synchrony is captured through a random-effects model taking into account multiple sources of variation and repeated measurements for each pair of children. Our findings indicate the presence of synchrony between participants during game play, which increases as they become more engaged in the game. These results are discussed in relation to personalized human-computer interaction and adaptive game environments.

Keywords

synchrony, speech, language, children, game design, engagement

1. INTRODUCTION

Interpersonal synchrony or entrainment refers to the coordination of behavioral patterns between interacting people [2]. Coordinated vocal, linguistic, and gestural behavior in adults can increase social connection [6], enhance rapport and affiliation [9], and facilitate task cooperation [14]. Similar benefits with respect to social behavior [8], repair [21], and attention [12] are found in infants and children. Specifically during children's game playing, synchronized behavior can promote closeness and social rapport [15].

An essential factor in successful game design is engagement [16]. In digital games and learning tools, engagement is related to enjoyment [10] and learning success [18]. Recently focus has shifted to detecting players' engagement and affect in order to enhance the gameplay experience through automatic adaptation [19]. Similar studies have demonstrated robots' use of engaging strategies for promoting the social aspects of human-robot interaction [20]. Other efforts have focused on automatic prediction of children's engagement through vocal [7] and visual [17] cues.

We explore children's synchrony during a two-party speech-driven, computer-based game and its relation to participants' engagement levels. We quantify children's verbal and acoustic behavior through descriptors of frequency, duration and prosody. Synchrony is modeled through random-effects models because of their ability to take into account a wide variety of correlation patterns [5]. Results indicate that children's verbal and acoustic patterns during game play are significantly correlated. Moreover, there is an interaction with level of engagement such that more engaged pairs show higher synchrony in word rate, speech loudness and fundamental frequency.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
INTERPERSONAL'15, November 13, 2015, Seattle, WA, USA.
© 2015 ACM. ISBN 978-1-4503-3986-5/15/11 ...$15.00.
DOI: http://dx.doi.org/10.1145/2823513.2823518.

2. DATA DESCRIPTION

2.1 The Game of Mole-Madness

"Mole Madness" is a speech-controlled interactive game similar to Super Mario Bros® video games, in which players move a "mole" character through its environment acquiring rewards and avoiding obstacles [13]. One child creates horizontal movement using the word "go" and the other creates vertical movement with "jump." Both coordinated turn-taking and overlapping speech are required for successful play.

2.2 Participants

Our data include 16 children (ages 6-10 years), 9 girls and 7 boys. Pairs were friends or siblings with an average age difference of 8 months. The recording equipment included two high definition cameras and two stereo microphones. The average duration of each game was 337.25 ± 56.47 seconds.

2.3 Engagement Annotation

Videos were split into 10-second intervals to preserve adequate context with low heterogeneity. Segments were duplicated and then modified to show only the child on the left or the child on the right. The resulting 442 segments were randomized and scored by each of three female coders experienced with young children. Coders were given a 7-point scale (ranging from 1 to 7) that asked them to make a concrete judgment regarding the child's willingness to continue playing the game (highest value) versus being ready to do something else (lowest value). More details on the annotation procedures can be found in [1].

For the purposes of this analysis, the engagement scores (ES) of both children were averaged over each segment. Since we are looking at coordination phenomena, this allows

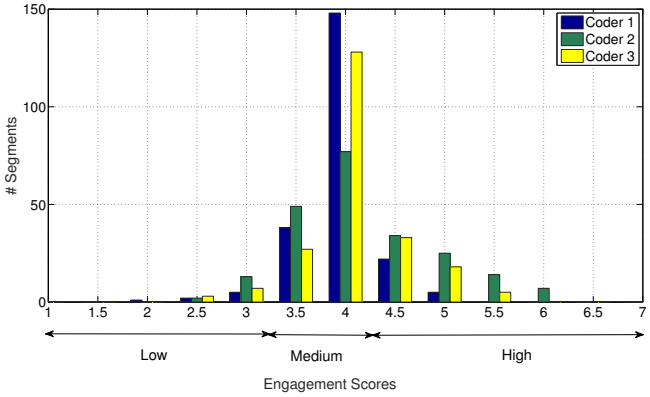


Figure 1: Distribution of engagement scores across coders 1, 2, and 3.

us to get an engagement approximation for the interacting pair along each 10-sec interval and further simplifies the statistical analysis. The resulting distribution (Fig. 1) led to the use of a 3-level categorization into low, medium, and high engagement with corresponding values within the intervals $[1, 3.5]$, $[4, 4.5]$, and $[5, 7]$, respectively. The median of the assigned ES appears to be around 4 for all coders, while the range seems to be more variable. Because of this variability, we treat each of the coders separately in our analysis.

3. FEATURE DESIGN

The highly linguistic nature of the game led us to the exploration of verbal and acoustic features that are likely to capture the variable facets of the interaction as it unfolds. Segmentation of "gos" and "jumps", referred as "keywords", was accomplished through a combination of automatic and hand-coding methods to ensure the accuracy of boundaries.

3.1 Verbal Features

Although the game's simple two-word task vocabulary makes it easily accessible to even very young children, we have found that its restricted expressivity gives rise to linguistic variations across all age groups. The most common variations involve vowel elongation or the rapid repetition of the keyword or initial phone. Motivated by these observations, verbal features include the number of keywords (denoted as "#Keywords") and the mean keyword duration ("Keyword Duration") per child within the 10-second engagement interval (Section 2.3). Similar timing features were explored in previous research [11].

3.2 Acoustic Features

Due to the observed variation of children's prosodic patterns, we further study speech loudness and fundamental frequency (F0). These features are extensively used in children's prosody-related studies [7]. Loudness is computed as the normalized intensity raised to the power of 0.3, while F0 is detected through an autocorrelation method, as implemented in openSMILE [4]. For each 10-second interval, we computed the mean loudness and F0 value of all non-overlapping keywords per child. These features are also intuitive and easily reproduced for adapting the behavior of a robot character to a child.

4. RANDOM-EFFECTS MODEL

Random-effects models (REM) provide a flexible approach to the analysis of variable sources of correlation and are extensively used to model repeated data structures that might violate independence assumptions [5]. Although there have been other attempts to quantify synchrony [3], the use of this type of statistical model reflects the exploratory nature of our analysis, since it yields easily interpretable results that can be translated to a child-robot interaction scenario.

Let X and Y be a pair of children and X_{ij} and Y_{ij} be a verbal or acoustic score for the j^{th} pair during the i^{th} 10-second interval. Note that a model captures either a verbal or an acoustic feature. Also let $ES_{L_{ij}}$, $ES_{M_{ij}}$, $ES_{H_{ij}} \in \{0, 1\}$ be the corresponding low, medium, and high engagement group, obtained by assigning the average engagement value across the two children to one engagement category for a given 10-sec segment (Section 2.3). The linear REM algorithm relates the verbal or acoustic score for one child Y_{ij} to the corresponding score of the other child X_{ij}, a pair-specific mean that reflects that reflects differences across pairs, and the assigned engagement category using the following equation

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \beta_{2j} ES_{L_{ij}} + \beta_{3j} ES_{M_{ij}} + \beta_{4j} ES_{H_{ij}} + r_{ij}$$

where r_{ij} is the residual term and β_{0j} can be written as the sum of the grand-mean verbal/acoustic score γ_{00} and the pair-specific mean u_{0j}

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

The relation of verbal or acoustic measures between the two children is estimated as a fixed-effect

$$\beta_{1j} = \gamma_{10}$$

The same assumption holds for the impact of the different engagement levels on verbal/acoustic measures

$$\beta_{2j} = \gamma_{20}, \quad \beta_{3j} = \gamma_{30}, \quad \beta_{4j} = \gamma_{40}$$

These can be summarized with the following equation

$$Y_{ij} = \left\{ \begin{array}{ll} \gamma_{10} X_{ij} + \gamma_{00} + u_{0j} + \gamma_{20} + r_{ij}, & \text{low ES} \\ \gamma_{10} X_{ij} + \gamma_{00} + u_{0j} + \gamma_{30} + r_{ij}, & \text{medium ES} \\ \gamma_{10} X_{ij} + \gamma_{00} + u_{0j} + \gamma_{40} + r_{ij}, & \text{high ES} \end{array} \right\}$$

Generalized REM are an extension of linear REM to model response variables from various distributions [5]. We use the Poisson REM to model the number of keyword occurrences because of the counting nature of the corresponding outcome variable, and linear REM for the remaining variables.

5. RESULTS

Results of the random-effects model are presented in Table 1. The rows corresponding to γ_{00} and γ_{10} indicate the (non-)significant effect of intercept and verbal/acoustic features of one speaker to the corresponding features of the other. The rows related to variables γ_{30} and γ_{40} depict the percent change of the medium and high engagement group compared to the low one for a unit increase. This relative association between categorical classes is a standard practice when presenting REM results [5], that is why the coefficient of the first group (γ_{20}) is omitted. The graphical representation of the predicted outcome variables with respect to the three engagement levels for Coder 2, is shown in Fig. 2. Similar plots for the other coders are omitted due to space.

Table 1: *Random-effects model estimates for predicting the verbal and acoustic features of one speaker (Spk1) based on the corresponding features from the second speaker (Spk2) and in relation to the annotated engagement score (ES) levels.*

Feature	REM Estimate	Coder 1	Coder 2	Coder 3
# Keywords Spk1	Intercept γ_{00}	1.36** (0.12)	1.36** (0.12)	1.26** (0.14)
	# Keywords Spk2 γ_{10}	0.06** (0.00)	0.05** (0.01)	0.04** (0.01)
	ES Medium γ_{30}	0.21** (0.07)	0.25** (0.07)	0.35** (0.09)
	ES High γ_{40}	0.08 (0.19)	0.53** (0.08)	0.82** (0.13)
Keyword Duration Spk1	Intercept γ_{00}	0.34** (0.08)	0.34** (0.09)	0.32** (0.10)
	Keyword Duration Spk2 γ_{10}	0.29* (0.12)	0.29* (0.13)	0.31* (0.13)
	ES Medium γ_{30}	0.00 (0.04)	0.00 (0.04)	0.01 (0.06)
	ES High γ_{40}	0.25* (0.13)	0.02 (0.05)	0.06 (0.08)
Loudness Spk1	Intercept γ_{00}	-0.04 (0.09)	0.38 (0.25)	0.33 (0.21)
	Loudness Spk2 γ_{10}	0.94** (0.04)	0.44** (0.10)	0.59** (0.10)
	ES Medium γ_{30}	0.13 (0.09)	0.17* (0.07)	0.08 (0.11)
	ES High γ_{40}	-0.03 (0.31)	0.47** (0.10)	0.18 (0.16)
F0 Spk1	Intercept γ_{00}	229.06** (29.15)	259.27** (27.43)	207.59** (29.71)
	F0 Spk2 γ_{10}	0.29** (0.07)	0.20** (0.08)	0.38** (0.08)
	ES Medium γ_{30}	47.19** (12.75)	45.96** (11.31)	35.45* (16.23)
	ES High γ_{40}	126.45** (43.19)	72.24** (15.40)	15.46 (23.93)

*$p < 0.05$, **$p < 0.01$, parentheses denote standard deviation

Results indicate that there is a significant association between children across the considered verbal and acoustic features, suggesting the presence of verbal and acoustic synchrony during game play. A significant effect of engagement with respect to the number of keywords exists across annotators in almost all cases, indicating the importance of the corresponding variable for the considered framework. Higher keyword duration is significantly associated with high engagement, as scored by Coder 1 but not by the other coders. Finally, F0 seems an important feature across all coders, while the effect of engagement on loudness is apparent only for Coder 2. Taken together these findings suggest that even people who are extremely familiar with young children may use different features, or the same features to different degrees, in judging children's engagement in an activity.

6. DISCUSSION

This paper discusses verbal and acoustic synchrony patterns between children, and their relation to engagement during game play interactions. Our results indicate that despite the inherently limited linguistic nature of the considered game, the children find ways to verbally manifest their engagement and associate to their interacting peer. The relevant features are intuitive and fairly easily reproducible in a statistical synthesis framework. Taking this into account, future work will explore ways to produce parameterized acoustic spaces for multiple engagement categories, with the goal of using the models to inform the behavior of a robot co-player that can measure, sustain and enhance a child's engagement. Child-robot interaction data that have been collected as a part of this study will further help us validate these hypotheses.

Although our results suggest the presence of synchrony between children, which is likely to increase during high engagement instances, we have not, to date, differentiated between game- and enjoyment-elicited synchrony. The first occurs during instances where the children have to cooperate in order to make the mole avoid an obstacle or get reward points. The latter happens when the children coordinate verbally and rhythmically as an end in itself, not because the mole's environment requires it. Data inspection has revealed several instances where synchrony is achieved despite the game environment, such as examples of social parity (e.g. one child says as many words as the other) and instances of synchrony for synchrony's sake (e.g. both children singing go/jump together). In our future work, we plan to delineate those through time-logged events of the game.

Synchrony refers to the behavioral covariation between people [2]. Although it is difficult to entirely capture and quantify children's behavior during this multifaceted interaction, our results indicate an association at the acoustic signal level and at the extent of verbal repetition. Future work will concentrate on expanding those to facial expression and body-language behavior.

7. CONCLUSIONS

We analyzed children's synchrony with respect to quantifiable facets of their verbal and acoustic behavior during an interactive, speech-controlled video game. Our results indicate that the considered aspects of behavior are associated between children and are further related to their engagement levels, as judged by human coders. Findings also suggest the presence of heterogeneity in the perceived features for judging engagement across the three annotators. This exploratory analysis helps to better understand children's behavior during game tasks. It further provides a foundation towards building socially-appropriate assistive robots that can detect and flexibly adapt to change in engagement patterns in order to build rapport and produce an enjoyable game experience.

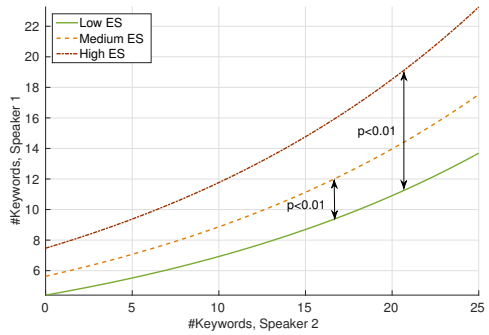
8. REFERENCES

[1] S. Al Moubayed and J. Lehman. Toward better understanding of engagement in multiparty spoken interaction with children. *ICMI*, 2015.

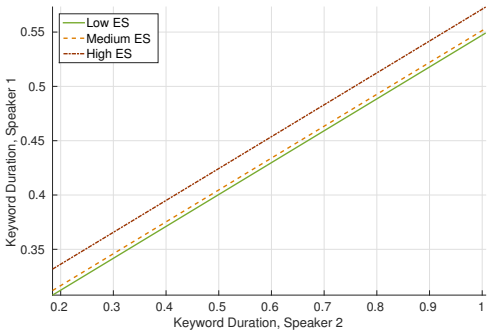
[2] E. Butler. Temporal interpersonal emotion systems: The "TIES" that form relationships. *Personality and Social Psychology Review*, 15(4):367–393, 2011.

[3] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE TAC*, 3(3):349–365, 2012.

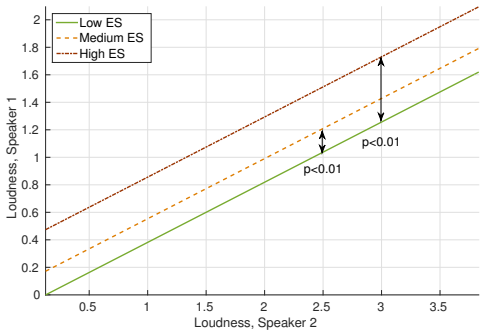
[4] F. Eyben, M. Wöllmer, and B. Schuller. openSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. ACM Multimedia*, pp. 1459–1462, 2010.



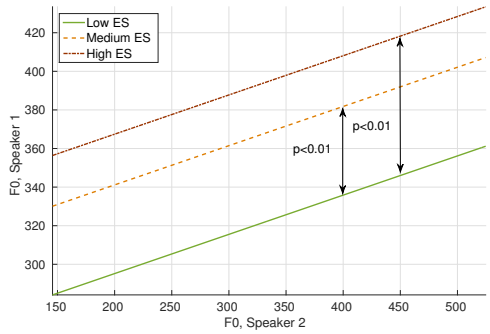
(a) #Keywords



(b) Keyword Duration



(c) Loudness



(d) F0

Figure 2: Graphical representation of predicted outcome values for low, medium and high engagement scores (ES) using random-effects model. ES were annotated by Coder 2.

[5] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

[6] A. Gonzales, J. Hancock, and J. Pennebaker. Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 2009.

[7] R. Gupta, C. Lee, S. Lee, and S. Narayanan. Assessment of a child's engagement using sequence model based features. In *Proc. WASSS*, 2013.

[8] A. Harrist and R. Waugh. Dyadic synchrony: Its structure and function in children's development. *Developmental Review*, 22(4):555–592, 2002.

[9] M. Hove and J. Risen. It's all in the timing: Interpersonal synchrony increases affiliation. *Social Cognition*, 27(6):949–960, 2009.

[10] A. Karimi and Y. Lim. Children, engagement and enjoyment in digital narrative. In *Proc. Ascilite*, pp. 475–483, 2010.

[11] M. Kehoe. Prosodic patterns in children's multisyllabic word productions. *Language, Speech, and Hearing Services in Schools*, 32(4):284–294, 2001.

[12] A. Khalil, V. Minces, G. McLoughlin, and A. Chiba. Group rhythmic synchrony and attention in children. *Frontiers in psychology*, 4, 2013.

[13] J. Lehman and S. Al Moubayed. Mole madness - a multi-child, fast-paced, speech-controlled game. In *AAAI Symposium on Turn-taking and Coordination in Human-Machine Interaction*, 2015.

[14] J. Manson, G. Bryant, M. Gervais, and M. Kline. Convergence of speech rate in conversation predicts cooperation. *Evolution and Human Behavior*, 34(6):419–426, 2013.

[15] T. Rabinowitch and A. Knafo-Noam. Synchronous rhythmic interaction enhances children's perceived similarity and closeness towards each other. *PloS One*, 10(4), 2015.

[16] J. Read, S. Mac Farlane, and C. Casey. Endurability, engagement and expectations: Measuring children's fun. In *Interaction design and children*, 2:1–23, 2002.

[17] J. Rehg, G. Abowd, A. Rozga, M. Romero, M. Clements, S. Sclaroff, I. Essa, O. Ousley, Y. Li, C. Kim, H. Rao, J. Kim, L. Presti, J. Zhang, D. Lantsman, J. Bidwell, and Z. Ye. Decoding children's social behavior. In *Proc. CVPR*, pp. 3414–3421, 2013.

[18] M. Ronimus, J. Kujala, A. Tolvanen, and H. Lyytinen. Children's engagement during digital game-based learning of reading: The effects of time, rewards, and challenge. *Computers & Education*, 71:237–246, 2014.

[19] N. Shaker, S. Asteriadis, G. N. Yannakakis, and K. Karpouzis. Fusing visual and behavioral cues for modeling user experience in games. *IEEE Transactions on Cybernetics*, 43(6):1519–1531, 2013.

[20] C. Sidner, C. Lee, C. Kidd, N. Lesh, and C. Rich. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1):140–164, 2005.

[21] S. Woltering, V. Lishak, B. Elliott, L. Ferraro, and I. Granic. Dyadic attunement and physiological synchrony during mother-child interactions: An exploratory study in children with and without externalizing behavior problems. *Journal of Psychopathology and Behavioral Assessment*, 1–10, 2015.