An Empirical Analysis of User Uncertainty in Problem-Solving Child-Machine Interactions

Matthew Black, Jeannette Chang, and Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory (SAIL) -- http://sail.usc.edu
University of Southern California, Los Angeles, CA, USA

matthepb@usc.edu, jeannetc@usc.edu, shri@sipi.usc.edu

ABSTRACT

With the widespread use of technologies directed towards children, child-machine interactions have become a topic of great interest. Computers must interpret relevant contextual user cues in order to provide a more natural interactive environment. Our focus in this paper is analyzing audio-visual user uncertainty cues using spontaneous conversations between a child and computer in a problem-solving setting. We hypothesize that we can predict when a child is uncertain in a given turn using a combination of acoustic, lexical, and visual gestural cues. First, we carefully annotated the audio-visual uncertainty cues. Next, we trained decision trees using leave-one-speaker-out cross-validation to find the more universal uncertainty cues across different children, attaining 0.494 kappa agreement with ground-truth uncertainty labels. Lastly, we trained decision trees using leave-one-turn-out cross-validation for each child to determine which cues had more intra-child predictive power and attained 0.555 kappa agreement. Both of these results were significantly higher than a voting baseline method but worse than average human kappa agreement of 0.744. We explain which annotated features produced the best results, so that future research can concentrate on automatically recognizing these uncertainty cues from the audio/video signal.

1. INTRODUCTION

The improvement of human-machine interaction (HMI) for young children has become an issue of great importance as youth have grown more comfortable with using new technologies. The potential contributions of automatic interactive systems for children are tremendous, especially in areas such as education and entertainment. Much HMI research has been dedicated to enhancing adult-machine interactions. However, differences in cognitive development and application-specific domain mismatches require that we examine children separately. For example, in one study involving one-word responses, adults were found to be much more consistent in expressing their uncertainty than children of ages 7-8 years. The authors hypothesized that adults were more concerned than children with self-presentation and "saving face" by clearly showing they were uncertain [8].

Studies specific to the needs of children in HMI have been relatively limited [1,2,9,13]. The focus of [1] was bolstering the performance of automatic tutorial systems for young adults. Analysis and detection of politeness and frustration states of children of ages 4-6 years in the Little CHildren's Interactive Media Project (CHIMP) database, the same corpus used in this study, was emphasized in [2,13]. This dataset consists of children playing an interactive computer game by conversing with a Wizard-of-Oz controlled character. Since the conversation is in a problem-solving setting, it is crucial that the computer automatically recognize when the child is uncertain. It has been shown that a person's state of uncertainty is displayed through communicative cues [4]. In an effort to enhance HMI, some have tried to identify the audiovisual cues most relevant to an adult's state of uncertainty [10,11]. Our goal is to show which audio-visual (lexical, acoustic, linguistic, visual/gestural) cues, annotated by humans, best predict when a child is uncertain, how well the cues generalize among different children, and the intra-child predictive power of the cues.

If machines were able to pick up on these subtle uncertainty cues of children, they could respond accordingly and provide the appropriate assistance. This could dramatically improve child-machine interaction by allowing for a more adaptive and personalized conversation. We hope this work will be a good first step and lay the groundwork for automatically sensing and processing when a child is uncertain directly from audio/video signals, an area of future research.

2. CORPUS

For this study, we used data from Little Children's Interactive Multimedia Project (CHIMP), collected by the Signal Analysis and Interpretation Laboratory at the University of Southern California in 2005. In the Little CHIMP experiments, 50 children of ages 4-6 each had a series of conversational interactions. They first had a brief introductory conversation with a human moderator. They then had a similar introductory session with an embodied computer character, "Josh," controlled in a Wizard-of-Oz setting. Next, they played an interactive computer game with Josh by answering a number of age-appropriate problem-solving questions (e.g., counting the number of objects on the screen, ordering a sequence of pictures, and summarizing stories). Josh provided feedback to the children throughout the game. The sessions concluded with debriefings with Josh and the human moderator. We analyzed the interactive game sessions from four boys and four girls, with each session approximately ten minutes long. The data collected included video with a front/side view of the child (Figure 1), as well as a single audio channel which included the speech of both the child and computer.

3. ANNOTATION SCHEME

We used the Anvil [6] software to annotate all eight video files, which allowed us to track the presence of multiple observable events in parallel (Figure 1). We marked the start and end times for the following events: 1) uncertainty, 2) speech transcriptions, 3) acoustic/linguistic events, and 4) visual/gestural events.
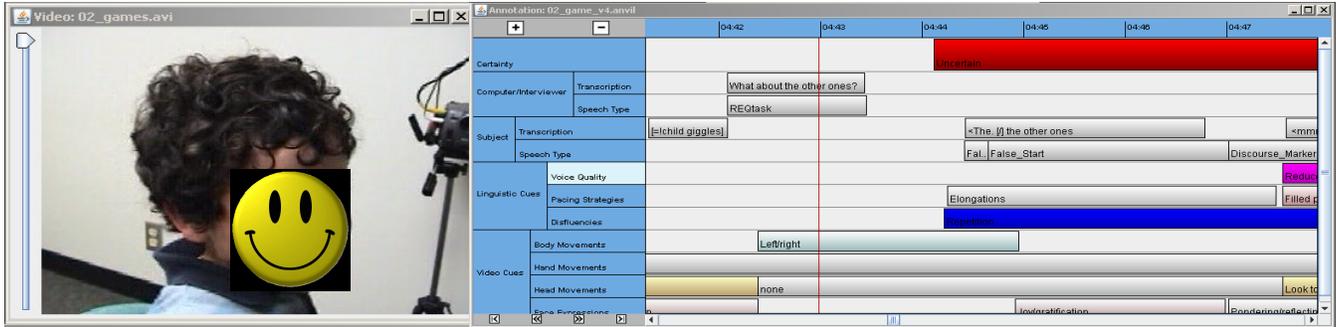
Figure 1: Child playing conversational game (face hidden for privacy reasons) with screenshot of Anvil [6] annotation scheme

3.1 Uncertainty

Three annotators marked segments during which the children appeared uncertain. This allowed us to quantify the degree of subjectivity for this measure through computation of agreement statistics. All annotators were provided with the following definition of uncertainty: "Uncertainty refers to when the user is in a state of doubt and lacks confidence. She/he appears confused and may be hesitant when prompted to provide an answer."

We initially defined agreement of uncertainty between annotators as any overlap in the time stamps. However, this resulted in a low mean pairwise annotator agreement of 62.40 percent, potentially due to a difference in annotation style. To account for this, we recomputed uncertainty agreement at the turn-level, where a turn is defined as the time between two consecutive instances of the computer speech. Analyzing uncertainty at the turn-level was sufficient for this research, since we were interested in finding the most relevant cues that were correlated with uncertainty i.e. occurring within the same turn. A turn was labeled as uncertain for a particular annotator if that annotator marked the child as uncertain at any time during the turn. Mean pairwise uncertainty agreement at the turn-level was 78.84 percent, significantly higher than when using the stricter agreement metric ($p < 0.01$).

Ground-truth uncertainty marks for each turn were calculated using a voting method: if at least two annotators marked the child as uncertain at any point during the turn, the ground-truth for that turn was the child being uncertain. Of the 271 turns in the corpus, 76 of the ground-truth uncertainty labels were marked as the child being uncertain (28.04 percent). The mean agreement between the ground-truth uncertainty labels and the three annotators was 89.42 percent (mean kappa agreement of 0.744). These ground-truth uncertainty labels served as the dependent variable for all classification experiments in this paper.

3.2 Uncertainty Cues

The independent variables were derived from the other three annotation streams, which served as potential cues to uncertainty. These cues were assumed to be inherently less subjective than labeling uncertainty since they were based on better-defined physical events. Therefore, they were only annotated by one person. For the children's speech, we recorded word-level transcriptions since some children might convey their uncertainty lexically (e.g., by saying, "I don't know"). Since in this study our focus was on deriving cues from uncertainty obtained from audio-visual information, we did not include analysis of the computer agent speech or the dialog context. Sections 3.2.1 and 3.2.2

describe the acoustic/linguistic and visual/gestural events, respectively, with Table 1 showing annotation statistics.

3.2.1 Acoustic and Linguistic Events

For the acoustic/linguistic events, we labeled three separate categories that we felt were potential uncertainty cues: voicing type, pacing strategies, and disfluencies. For each of these categories, we defined a discrete subset of options. Voicing type could be categorized as: strong, reduced, whispered, questioned, or the annotator could leave the category blank in cases when the voicing type was perceived as normal. Here, *strong* refers to the child talking in a loud voice, *reduced* refers to a quiet voice, *whispered* corresponds to a whispered voice, and *questioned* meant the child used a question intonation. We chose to mark these voicing type cues since we felt the use of a strong voice might indicate certainty and reduced, whispered, and question intonations may be correlated with the child being uncertain.

Pacing strategies were marked if the child was "buying time." There were two distinct types: pauses and elongations. Pauses were used to mark silence that was longer than two seconds, and elongations were marked if the child lengthened a syllable or phone to disrupt the flow of the word pronunciation. These features were used since we hypothesized the child might pause or elongate if uncertain about an answer.

Five types of disfluencies were explicitly marked: fillers, repetitions, partial-word repetitions, repairs, and false-starts. Fillers included when the child said "uhhh" or "ummm." Repetitions were marked in cases where children repeated whole words or phrases, while partial-word repetitions were marked if only syllables or phones within a word were repeated. Repairs were marked when the child self-corrected, and false-starts were marked when the child began a sentence or utterance, stopped, and began a new one. We marked these disfluencies since it has been shown that some speakers' disfluency rates fluctuate as a function of cognitive planning load [3].

3.2.2 Visual and Gestural Events

For the visual and gestural events, we marked the following: body movements, hand movements, head movements, and facial expressions. These events were marked since we figured there may be some type of visual indication that the child is uncertain. Discrete label options within the body movements included leaning close to the computer screen, leaning to the left/right, leaning back in the chair, and slumping. Hand movements included pointing to the computer screen, pointing elsewhere, raising either hand, and miscellaneous hand movements. Head

movements included shaking yes, shaking no, other head shakes, looking up, looking down, and looking to the side. Facial expressions were discretized into the following: happy, angry, helpless, thoughtful, and surprised. While the facial expressions may be more subjective than marking head movements, annotating with this discrete set of facial expressions was far less tedious than marking, for example, eyebrow movements.

Table 1. Total number of times each annotation stream/label was marked for the eight children in the corpus

Annotation Stream (Total #)	Annotation Label (Total #)
Acoustic: Voicing Type (174)	Reduced (62), Strong (70), Whispered (15), Question (27)
Linguistic: Pacing Strategies (334)	Pause (291), Elongation (43)
Linguistic: Disfluencies (446)	Filler (230), Repetition(82), Partial-word Repetition (21), Repair (41), False Start (72)
Visual: Body Movements (35)	Lean back (28), Left/Right (7)
Visual: Hand Movements (225)	Point Screen (3), Point Self (32), Point Else (68), Raise (24), Meaningless (98)
Visual: Head Movements (214)	Look Up (27), Look Down (26), Look Side (107), Shake No (16), Shake Yes (22), Shake Else (16)
Visual: Face Expressions (347)	Happy (72), Angry (41), Helpless (48), Surprised (20), Thoughtful (166)

4. FEATURE EXTRACTION

We extracted features from the human annotations at the turn-level to determine whether a child was uncertain in a given turn. For the word-level transcriptions, we used unigram and bigram word counts in a bag-of-words feature representation, since these features performed better than word- and character-level language models and have been used successfully in many text classification studies (such as [7]). For the acoustic/linguistic and visual/gestural annotation streams, we extracted two features for each label: the number of times the label was used in a turn (referred to as *count* in this paper) and the percentage of the turn length marked as the label (referred to as *percentage*). Since many of the labels were only sparsely used in the data (Table 1), we also extracted the count and percentage for each annotation category. For example, rather than just extracting the count and percentage of the label, "shaking head yes," we also extracted the count and percentage of, "head movements." One final feature chosen was the length of the first pause in the turn, which was included since many of the children would pause initially if asked a difficult question. Many more features could be envisioned and used, but we limited ourselves to these features for this study.

5. CLASSIFICATION OF UNCERTAINTY

The ground-truth uncertainty labels served as the dependent variable in these experiments, with features described in Section 4 serving as the independent variables. We used the alternating decision tree [5] in WEKA [12] for this binary classification problem. Decision trees were used since they are known for their good performance on tasks involving both discrete (count) and continuous (percentage) features in a classification framework and have a built-in feature selection process. Alternating decision trees were used over the more traditional C4.5 decision trees

because they performed better in initial cross-validation experiments, perhaps due to the use of boosting in the alternating decision tree training algorithm [5].

We performed two classification experiments by partitioning the data in two ways. We first trained the decision trees using leave-one-*speaker*-out cross-validation. We called this the *inter-child* experiment, since it was intended to find the more universal cues that predicted uncertainty across different children. Next, we partitioned the data using leave-one-*turn*-out cross-validation for each child separately. We called this the *intra-child* experiment since we hoped to find which uncertainty cues were child-specific and could model the idiosyncratic uncertainty behaviors of each child. We were initially worried about data sparsity issues for this experiment, but 30 to 40 turns were sufficient to get statistically significant results. For both experiments, we tuned the decision trees' parameters using ten-fold cross-validation on the training set to maximize kappa agreement with the ground-truth uncertainty labels. Finally, we used these optimized decision trees on the test data and quantified classification performance using two metrics: percent agreement and kappa agreement.

6. RESULTS & DISCUSSION

Table 2 shows the classification agreement statistics using different feature subsets. For the inter-child experiment, only the *voicing type* features by themselves attained a percent agreement that was significantly higher ($p<0.1$) than the voting baseline; the "reduced voice percentage" and "question count" labels appeared in all of these decision trees. The lexical features were next best, with the word, "I," being the best lexical *uncertainty* cue (in the context of saying, "I don't know" or "I can't") and the word, "because," being the best lexical *certainty* cue. Combining all pairs and triplets of feature subsets for the inter-child experiment, we reached a percent agreement of 81.55 and kappa agreement of 0.494 with the *voicing type* and *lexical* features together. This implies that these features are good general cues towards uncertainty and should be considered when trying to recognize uncertainty in different children using general children's models.

For the intra-child experiments, the *pacing strategies* features produced the best results, with the decision trees using the "first pause length" and count and percentage of pauses the most. Using the *disfluency* features by themselves also had significantly higher percent agreement over the baseline system. The decision trees for the *disfluency* features greatly differed from one another, suggesting that children have their own unique way of expressing uncertainty with disfluencies. For the intra-child experiments, the combination of the *pacing strategies*, *voicing type*, and *facial expressions* features attained the highest percent agreement of 83.76 and kappa agreement of 0.555. This implies that this combination of cues can predict uncertainty in children, but child-specific models may need to be trained.

Overall, the intra-child agreement statistics were higher than the inter-child agreement statistics, probably due to the large variation in the features across different children. However, the inter-child experiment may have also suffered because of the small number of annotated children in the corpus; we may need more than eight children to get good generalizable results. The best combination of features for both experiments were still significantly worse when compared to average human agreement (both $p<0.05$). This performance gap could be due to the inherent difficulty of using

objective measures to predict a subjective measure. It could also be due to the turn-level analysis and choice of features, which resulted in the loss of potentially important temporal and cause-and-effect information. In addition, we assumed the turns are independent in our analysis, which is not realistic in this continuous interactive setting.

Table 2. Classification agreement statistics when training alternating decision trees using different subsets of features. Bold red numbers mean performance was significantly better than baseline (p<0.1).

Feature Subset	Inter-Child		Intra-Child	
	Percent	Kappa	Percent	Kappa
Baseline (Voting)	71.96	0.000	71.96	0.000
Lexical (Bag-of-Words)	76.39	0.304	75.28	0.206
Linguistic: Disfluencies	71.22	0.197	*78.23*	0.384
Linguistic: Pacing Strategies	72.32	0.266	*82.66*	0.518
Acoustic: Voicing Type	*80.81*	0.463	76.01	0.267
Visual: Body Movements	68.27	0.000	72.32	0.082
Visual: Facial Expressions	71.96	0.222	75.28	0.313
Visual: Hand Movements	70.48	0.132	72.69	0.145
Visual: Head Movements	69.37	0.093	76.75	0.336
Best: Voicing, Lexical	*81.55*	0.494		
Best: Pacing, Voicing, Facial			*83.76*	0.555
Average Human Agreement	89.42	0.744	89.42	0.744

7. SUMMARY & FUTURE WORK

We found that *voicing type* features (especially the use of a quiet voice and question intonation) are the best global cues for uncertainty in children, and *pacing strategies* (especially pauses) and *disfluency* cues are better at predicting child-specific uncertainty trends. We determined these findings by training alternating decision trees with features extracted at the turn-level from human annotations on different partitions of the data. Average human kappa agreement that a child was uncertain in a turn was 0.744, and we achieved kappa agreements of 0.494 and 0.555 for the inter-child and intra-child experiments, respectively.

This research established how difficult it is to predict uncertainty in young children in this continuous interactive problem-solving conversational setting. In the future, we will work to resolve some of the limitations discussed earlier. For example, we hope to make use of information embedded in turn inter-dependencies rather than taking each turn as an independent occurrence. Because our corpus features a continuous setting, we believe that looking at context will provide further insight (such as cause-and-effect chain events) into understanding the user's affective state. We plan on using Markov models to incorporate information from previous turns to better predict the child's uncertainty state in a given turn. Taking advantage of the continuous nature of the Little CHIMP data would be a novel approach to user uncertainty detection, since most previous corpora have largely been limited to isolated utterances. In addition, we would like to explore features that address temporal aspects of our data. Although we included percentage features that took into account the duration of a given cue in relation to the turn length, it will be interesting to include features that identify during what part of a turn a cue occurs. Moreover, an alternative method of establishing ground-truth uncertainty could uncover links between timing issues and user uncertainty, which the turn-based analysis may not have addressed thoroughly.

Another approach we believe may yield more promising results is differentiating between degrees of uncertainty rather than simplifying classification to the binary decision of uncertain versus certain. It is likely that turns classified as strong cases of user uncertainty will help us identify more robust uncertainty cues than on-the-border cases. We hope to eliminate the "noise" from close-call instances of user uncertainty and thus fine tune our decision trees. The next step in this research is to begin automatically identifying the most relevant cues to uncertainty directly from the audio/video signals. Our long-term goal is to incorporate this research into real-time conversational programs that provide an educational game experience by adapting to the situation and needs of the child.

8. ACKNOWLEDGMENTS

9. REFERENCES

[1] Ai, H., Litman, D. J., Forbes-Riley, K., Rotaru, M., Tetreault, J., Purandare, A. 2006. Using system and user performance features to improve emotion detection in spoken tutoring dialogs. In Proc. of InterSpeech (Pittsburgh, USA, 2006).

[2] Arunachalam, S., Gould, D., Andersen, E., Byrd, D., and Narayanan, S. 2001. Politeness and frustration language in child-machine interactions. In Proc. of Eurospeech (Aalborg, Denmark, 2001).

[3] Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., and Brennan, S. E. 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. Language and Speech 44-2 (2001), 123-147.

[4] Brennan, S. E. and Williams, M. 1995. The feeling of another's knowing: prosody and filled pauses as cues to listeners about the metacognitive state of speakers. Journal of Memory and Language 34 (1995), 383-398.

[5] Freund, Y. and Mason, L. 1999. The alternating decision tree learning algorithm. Proc. Machine Learning (Bled, Slovenia, 1999).

[6] Kipp, M. 2004. Gesture generation by imitation - from human behavior to computer character animation. dissertation.com (Boca Raton, Florida, Dec. 2004).

[7] Koller, D. and Sahami, M. 1997. Hierarchically classifying documents using very few words. Proc. of International Conference on Machine Learning (Nashville, Tennessee, USA, 1997).

[8] Krahmer, E. and Swerts, M. 2004. Signaling and detecting uncertainty in audiovisual speech by children and adults. In Proc. of Interspeech (Jeju Island, Korea, 2004).

[9] Narayanan, S. and Potamianos, A. 2002. Creating conversational interfaces for children. IEEE Transactions on Speech and Audio Processing vol. 10-2 (Feb. 2002), 65-78.

[10] Smith, V. L. and Clark, H. H. 1993. On the course of answering questions. Journal of Memory and Language 32 (Feb. 1993), 25-38.

[11] Swerts, M., Krahmer, E., Barkhuysen, P., van de Laar, L. 2003. Audiovisual cues to uncertainty. ISCA workshop on error handling in spoken dialog systems (Chateau-d'Oex, Switzerland, 2003).

[12] Witten, I. H. and Frank, E. 2005. Data Mining: Practical machine tools and techniques. 2nd ed., Morgan Kaufmann, San Francisco.

[13] Yildirim, S., Lee, C. M., Lee, S., Potamianos, A., and Narayanan, S. 2005. Detecting politeness and frustration state of a child in a conversational computer game. In Proc. of Eurospeech (Lisbon, Portugal, 2005).