Detecting Politeness and Frustration State of a Child in a Conversational Computer Game

Serdar Yildirim, Chul Min Lee, Sungbok Lee, Alexandros Potamianos[*]*, Shrikanth Narayanan*

Speech Analysis and Interpretation Laboratory, Integrated Media Systems Center
Department of Electrical Engineering, Viterbi School of Engineering
University of Southern California, Los Angeles, CA 90089, USA
[*] Department of ECE, Technical University of Crete, Chania 73100, Greece
http://sail.usc.edu
yildirim@sipi.usc.edu

10.21437/Interspeech.2005-700

Abstract

In this study, we investigate *politeness* and *frustration* behavior of children during their spoken interaction with computer characters in a game. We focus on automatically detecting *frustrated*, *polite* and *neutral* attitudes from the child's speech (acoustic and language) communication cues and study their differences as a function of age and gender. The study is based on a Wizard-of-Oz dialog corpus of 103 children playing a voice activated computer game. Statistical analysis revealed that there was a significant gender effect on politeness with girls in this data exhibiting more explicit politeness markers. The analysis also showed that there is a positive correlation between frustration and the number of dialog turns reflecting the fact that longer time spent solving the puzzle of the game led to a more frustrated child. By combining acoustic and language cues for the task of automatic detection of politeness and frustration, we obtain average accuracy of 84.7% and 71.3%, respectively, by using age dependent models and 85% and 72%, respectively, for gender dependent models.

1. Introduction

The need for automatic recognition of user's emotion within a spoken dialog system framework has received increased attention in the past few years. The primary motivation is to make the machine interactions natural and more responsive to a user's behavior. However, a majority of the previous work on emotion recognition primarily targeted adult users even though children are one of the potential beneficiaries of computers with spoken interfaces such as for educational applications. Recognizing a child's emotional state during an interaction with a computer may help us to build interfaces that are more tuned to the child's needs. For instance, an interface that can mirror a child's politeness, and that can respond to frustration, in meaningful ways can help increase the naturalness of the interaction. This paper focuses on analyzing and detecting politeness and frustration behavior of children from their spoken language cues.

Automatic emotion recognition from speech is a challenging research problem in various respects including discerning the most appropriate signal features and classification methods. It is well acknowledged that the greater variability in the acoustic and linguistic characteristics of children's speech, and changes in those parameters with age and gender, pose significant challenges for building spoken dialog applications for children [8]. Hence one could expect the problem of emotion recognition from children's speech to face similar challenges. In this paper, we analyze certain emotional/attitudinal behavior of children during their spoken interaction with a computer game as a function of their age and gender as well as the task factors such as number of user turns and success in problem solving.

A preliminary study of politeness and frustration language in child-machine interactions was reported for different age groups by [2]. Their study indicated that younger children use less overt politeness markers and express more frustration compared to older children. The study was, however, limited to the linguistic characteristics of politness and frustration. In this paper, we not only extend the analysis to include both acoustic and language information but also address the problem of automatic detection of *frustrated* and *polite* states in the child-computer interaction.

Many previous efforts have addressed emotion recognition by employing pattern recognition techniques using speech acoustic features [3, 4]. For example, in [6], an accuracy of 80.53% was reported in the recognition of student emotional state in a corpus of human-human spoken tutoring dialogs using only acoustic features. In addition to acoustic features, the effect of language information on emotion classification has been reported in [1, 5]. Lee et al. [5] proposed the notion of emotional salience, i.e., mutual information between a specific word and an emotion class, to identify emotional words in speech utterance for detecting negative emotion. By adding language information to the acoustic features, they reported a relative improvement of 46%. Similarly, in [7], it was shown that the use of speech and language features for predicting student emotions in computer-human tutoring dialogues improved the accuracy of the system. Likewise, Zhang et al. [10] reported promising results in the combined use of acoustic, spectral, and language information in detecting confidence, puzzlement, and hesitation in their child-machine dialog task. On the other hand, the results of Ang et al. [1] where language model features measured from class-based trigram model were added to prosodic decision trees indicated that their language model features were poor predictors of frustration. In our work, we extend the notion of emotional salience in language by calculating the mutual information between word pairs and emotion classes. We use this information in conjunction with a variety of acoustic features.

This paper is organized as follows. Section 2 describes and analyzes the speech database we used. Section 3 explains the feature extraction procedure and experimental setup. Results and discussion are given in Section 4. Section 5 concludes the

female	male	7-9 y/o	10-11 y/o	12-14 y/o
48	55	38	35	30

Table 1: Distribution of subjects considered in this study according to their age and gender.

paper.

2. Speech Database

The speech data used in this paper, the Children's Interactive Multimedia Project (ChIMP) database, came from a study on child-machine interactions in a game setting [8][9]. The database contains spoken dialog interactions from 160 boys and girls, eight to fourteen years of age. A Wizard of Oz (WoZ) technique was used for data collection that resulted in over 50000 utterances. The task is to play "Where in the USA is Carmen Sandiego?", an interactive computer game using speech. The goal of the game was to identify and arrest a cartoon criminal. During the game, the child had to interact with several cartoon characters to obtain clues about the suspect. Most children played the game twice. Further details may be found in [8].

To date, utterances from 103 subjects out of 160 total subjects were labeled into one of three categories, *neutral*, *polite*, and *frustrated* by two native speakers of English. The agreement between the two annotators in terms of Kappa statistics is 0.63. Utterances that both annotators agreed on (over 15000) were considered in this study. The distribution of number of subjects according to age and gender is given in Table 1 and the distributions of the emotional categories are given in Table 2.

	Total	Neutral	Polite	Frustrated
7-9 y/o	37%	69%	17%	14%
10-11 y/o	35%	73%	20%	7%
12-14 y/o	28%	68%	16%	16%
Male	54%	72%	15%	13%
Female	46%	70%	20%	10%

Table 2: Emotional data distribution (%) for each age group and gender.

2.1. Analysis of emotional/attitudinal behavior of children

In order to determine relations between emotional categories and the number of dialog turns, we calculated the Pearson correlation coefficients using the SPSS statistical software package. Results showed that there is a positive correlation (r=0.259, p=0.001) between the number of total dialog turns and child frustration, indicating that as number of total dialog turn increases, the likelihood for frustration also increases. Simple factorial analysis (ANOVA) showed that the effect of age on the normalized frustrated user turns (i.e., the number of frustrated user turns in one game divided by the total number of turn in that game), is significant [F=4.791, $p<0.01$]. Multiple comparisons test indicates that 10-11 y/o group shows significantly less frustration than other age groups. Also the effect of age on the number of dialog turns is significant [F=8.05, $p<0.01$] indicating that as age increases the length of dialog decreases. The normalized number of frustrated and polite user turns were compared by task success (win/loss) and gender. The female children express more politeness and less frustration than the male children (20% vs 15%, 10% vs 13%, respectively), and also the games that ended with lost contain more frustrated user turns than that of the games ended with win (0.7% vs 0.5%). Results also showed that there is a negative correlation between the numbers of frustrated and polite user turns in a dialog (Kendall's tau-b=-2.87 $p<0.01$).

3. Methodology

3.1. Feature Extraction

Both acoustic and language information are used for predicting child's emotions. The feature set for acoustic information was derived from the F0 contour, energy and duration that were obtained directly from the speech signal. Language features were obtained by calculating mutual information between word pairs and emotion classes.

3.1.1. Acoustic Information

As acoustic features, we used 16 different parameters comprising utterance level statistics corresponding to F0 (fundamental frequency), energy, and duration. The ESPS Xwaves+ toolkit was used for the extraction of features. Details of parameters used are summarized below.

F0: Mean, median, standard deviation, maximum, and minimum.

Energy: Mean, median, standard deviation, maximum, and minimum.

Duration: utterance duration, average voiced and unvoiced duration, inter-word silence duration, longest voiced portion duration, speaking rate.

3.1.2. Language Information

Another source of information for emotion classification relates to the language usage in the spoken utterances. We obtained the emotional word pairs in this data by automatically calculating *emotional salience* of the word bigrams in the data corpus by following a similar approach suggested in [5] for word unigrams. Emotional salience of a word pair is a measure of how much information the word pair provides about the emotion category. Let wp denote the word pair in the database and $E = \{e_1, e_2, ..., e_k\}$ denote the emotional space, then the salience of word pair wp is:

$$sal(wp) = \sum_{j=1}^{k} P(e_j|wp) \log \frac{P(e_j|wp)}{P(e_j)} \quad (1)$$

The emotionally salient word pairs with respect to emotion category are the ones that have greater salience values.

After identifying the emotionally salient word pairs, we calculated the language features at the utterance level. Let $W = \{wp_1, wp_2, ..., wp_l\}$ be the word pairs of a utterence where l is the number of the word pairs in the utterance. We calculate the lexical features of this utterance for a given emotion as follows,

$$a_k = w_k + \sum_{m=1}^{l} I_m w_{mk} \quad (2)$$

where I_m denotes indicator, which has either 0 or 1 representing either a word pair matched to a salient word pair or not, w_{mk} denotes connection weight, and w_k is the bias. We can define

the connection weights w_{mk}, and bias w_k as follows:

$$w_{mk} = i(wp_m, e_k) = \log \frac{P(e_k|wp_m)}{P(e_k)} \quad (3)$$

$$w_k = \log P(e_k) \quad (4)$$

3.2. Experimental Setup

We used a linear discriminant classifier (LDC) which assumes that each class has a Gaussian probability density with common covariance. The task was to identify the emotional state of spoken utterances. Both feature level and decision level fusions were considered. In the feature level integration, the acoustic and language features were combined and a single classifier was used. In the decision level integration, seperate classifier was used for each information source and the outputs were combined using a product rule in which posterior probabilities are multiplied and the maximum is selected. The reason for using the product rule for decision level integration is that it outperformed the other decision fusion rules we investigated, such as maximum, minimum and averaging, for the task at hand. The performances of the classifiers are presented in terms of confusion matrix and overall accuracy (percentage of correctly classified utterances).

To investigate the gender- and age-dependencies in classification, each age group and gender data were considered separately. Since there is a large skew in our class sizes for each age group and gender, first, neutral data for each age group and gender was randomly divided into disjoint sets to have equal class priors [1, 5] resulting in different data sets for each age group and gender. Then for each data set, the performance was evaluated by tenfold cross-validation. The final evaluation metrics were calculated by combining results from each data set. Since politeness represents an attitude in speaking style while frustration reflects the user's emotional state, the detection of politeness and frustration was conducted separately.

4. Results and Discussion

Detection of politeness and frustration based on acoustic features, language features, and the feature level fusion method as a function of age and gender are reported and compared with the baseline accuracy where the classifier was built and evaluated using all data (without separating data for each gender and age groups). Baseline results are given in Table 3. Decision level information fusion scores were not given because for all cases, the feature level integration gave better results than that of the decision level integration for the task at hand.

	Politeness	Frustration
Acoustic	65	62
Language	81	64
Feature Level	84	70

Table 3: Classification accuracy (%) for the baseline.

4.1. Age-dependent Performance Evaluation

4.1.1. Detecting Politeness

Table 4 shows the overall classification accuracy for politeness for each age group. Best performance was achieved when acoustic and language information were fused at the feature level for all age groups. Notice that language information has more discriminative power than the acoustic cues for detecting politeness. This may be due to the fact that the word usage variations for politeness are somewhat limited in this data; for example, it was marked by few highly frequent phrases such as *please, thank you, excuse me*. Therefore contribution of lexical information increases so as to improve the performance in the classification yielding relative improvements of $\sim$30% over using only acoustic information, for all ages.

	7-9 y/o	10-11 y/o	12-14 y/o	Ave.
Acoustic	65	67	65	65.7
Language	82	83	77	81
Feature Level	83	87	84	84.7

Table 4: Classification accuracy (%) for the politeness detection task for each age group. Ave.: Average performance across age groups.

4.1.2. Detecting Frustration

	7-9 y/o	10-11 y/o	12-14 y/o	Ave.
Acoustic	65	68	62	65
Language	67	59	62	62.7
Feature Level	74	70	70	71.3

Table 5: Classification accuracy (%) for the frustration detection task for each age group. Ave.: Average performance across age groups.

Table 5 displays the overall accuracy results for frustration for each age group. Notice that the relative contribution of language information was significantly lower when compared to that observed in the politeness detection. The underlying reason is that there is greater overlap between word choices for neutral and frustration emotion states. The best performance was achieved when acoustic and language information were combined at the feature level. Relative improvements of 13.85%, 2.94%, and 12.90% were achieved over using only acoustic information for 7-9 y/o, 10-11 y/o and 12-14 y/o, respectively.

4.2. Gender-dependent Performance Evaluation

4.2.1. Detecting Politeness

	Female	Male	Ave.
Acoustic	65	62	63.5
Language	83	82	82.5
Feature Level	86	84	85

Table 6: Classification accuracy (%) of the politeness detection task for each gender. (Ave.: Average performance across gender.)

Table 6 shows the overall classification accuracy of politeness for each gender. The best performance was achieved when acoustic and language information were fused at the feature level. Notice that, acoustic cues were more informative for female than that of male children. Relative improvements of 32.31% and 35.48% were achieved by combining information cues at the feature level over using only acoustic cues.

Table 7 shows the confusion matrices of the classifiers when the acoustic and language cues were fused at the feature level.

	Female		Male	
	Neutral	Polite	Neutral	Polite
Neutral	83.81	16.19	78.64	21.36
Polite	12.04	87.96	11.57	88.43

Table 7: Confusion matrix for the politeness detection task based on the feature-level integration for each gender.

4.2.2. Detecting Frustration

	Female	Male	Ave.
Acoustic	64	65	64.5
Language	63	62	62.5
Feature Level	70	74	72

Table 8: Classification accuracy (%) of the frustration detection task for each gender. (Ave.: Average performance across gender.)

Table 8 shows the overall classification accuracy of frustration for each gender. Best performance was achieved when the acoustic and language information were fused at the feature level. Notice that, the acoustic cues were more informative than language information for both male and female children. Relative improvements of 9.37% and 13.84% were achieved by combining information cues over using only acoustic cues.

	Female		Male	
	Neutral	Frustration	Neutral	Frustration
Neutral	71.91	28.09	77.02	22.98
Frustration	33.04	66.97	29.05	70.95

Table 9: Confusion matrix of the frustration detection task based on feature-level integration for each gender. Values are given in percentage.

Table 9 shows the confusion matrices of the classifiers when the acoustic and language cues were fused at the feature level. As can be observed, performance of detecting frustration for male children is better than that of female children.

5. Conclusion

In this work, the detection of politeness and frustration emotions in child-computer spoken dialogs was examined using acoustic and language information. For all age groups and gender, combining the two information sources improves the performance of classifiers. Language information was a major factor in detecting *politeness*, but not in *frustration*. The recognition results showed that the child's age and gender affect the classification performance. Realization of emotions differs with age and gender. Performance of the female model is better than the baseline for politeness whereas the male model performed better for frustration. Perhaps, this may be a socia-linguistic effect, where politeness is more explicitly marked in the speech and language of the girls. The results for frustration detection were better for the male model. Again, one could speculate that boys more directly express frustration in their speech and language. There is a positive correlation between frustration and the number of total dialog turns, indicating that the number of frustrated turns increase as the number of total dialog turns increase. Statistical analysis revealed that the effect of age on emotion is significant, 10-11 y/o showed less frustration than the other age groups. The game scenario and the degree of difficulty may fit best for this age group, or it may be an artifact of this data, not a true age-dependent trend. However, the exact reason is unknown. There are several issues that must be further explored in the future. One challenging issue is to localize emotionally hot spots or frustrated user response by dynamically tracking the changes associated with acoustic and language variables as dialogue advances. Other sources of information that can be useful for that purpose are linguistic syntax, register and disfluency. The database used in this study also includes visual information which is another source of information to improve the overall performance. These are the directions of our ongoing efforts.

6. Acknowledgements

We would like to thank our colleagues in the Emotion Research Group of the Speech Analysis and Interpretation Lab (SAIL). This research was partially supported by Integrated Media Systems Center (IMSC) and NSF ERC under cooperative agreement No. EEC-9529152 and an NSF Career award.

7. References

[1] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proc. of ICSLP*, Denver, CO, 2001.

[2] S. Arunachalam, D. Gould, E. Andersen, D. Byrd, and S. Narayanan, "Politeness and frustration language in child-machine interactions," in *Proc. Eurospeech*, 2001, pp. 2675–2678.

[3] A. Batliner, K. Fischer, R. Huber, J. Spiker, and E. Noth, "Desperately seeking emotions: Actors, wizards, and human beings," in *Proc. ISCA Workshop on Speech and Emotion*, Belfast, 2000, pp. 195–200.

[4] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *ICSLP '96*, Philadelphia, PA, 1996.

[5] C. M. Lee and S. Narayanan, "Towards detecting emotions in spoken dialogs," *IEEE Trans. on Speech and Audio Processing*, 13(2), 293-303, 2005.

[6] D. J. Litman and K.Forbes-Riley, "Recognizing emotions from student speech in tutoring dialogues," *in Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, st. Thomas, Virgin Islands, December 2003.

[7] D. J. Litman and K.Forbes-Riley, "Predicting student emotions in computer-human tutoring dialogues," *in Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, Spain, July 2004.

[8] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Trans. On Speech and Audio Processing*, vol. 10(2), pp. 65–78, 2002.

[9] A. Potamianos and S. Narayanan, "Spoken dialog systems for children," in *ICASSP 98*, Seattle, WA, 2001, pp. 197–200.

[10] T. Zhang, M. Hasegawa-Johnson, and S. E. Levinson, "Children's emotion in an intelligent tutoring scenario," *in Proceedings of ICSLP*, Korea 2004.