Detecting Emotional State of a Child in a Conversational Computer Game[☆]

Serdar Yildirim[*,a], Shrikanth Narayanan[b], Alexandros Potamianos[c]

[a]*Computer Engineering Department, Mustafa Kemal University, Antakya, 31040, Turkey*
[b]*Signal Analysis and Interpretation Laboratory (SAIL), Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089, USA*
[c]*Department of ECE, Technical University of Crete, Chania 73100, Greece*

Abstract

The automatic recognition of user's communicative style within a spoken dialog system framework, including the affective aspects, has received increased attention in the past few years. For dialog systems, it is important to know not only what was said but also how something was communicated, so that the system can engage the user in a richer and more natural interaction. This paper addresses the problem of automatically detecting "frustration", "politeness", and "neutral" attitudes from a child's speech communication cues, elicited in spontaneous dialog interactions with computer characters. Several information sources such as acoustic, lexical, and contextual features, as well as, their combinations are used for this purpose. The study is based on a Wizard-of-Oz dialog corpus of 103 children, 7-14 years of age, playing a voice activated computer game. Three way classification experiments, as well as, pairwise classification between polite vs. others and frustrated vs. others were performed. Experimental results show that lexical information has more discriminative power than acoustic and contextual cues for detection of politeness, whereas context and acoustic features perform best for frustration detection. Furthermore, the fusion of acoustic, lexical and contextual information provided significantly better classification results. Results also showed that classification performance varies with age and gender. Specifically, for the "politeness" detection task, higher classification accuracy was achieved for females and 10-11 years-olds, compared to males and other age groups respectively.

Key words: emotion recognition, spoken dialog systems, children speech, spontaneous speech, natural emotions, child-computer interaction, feature extraction

1. Introduction

An emerging trend in Human-Computer Interaction (HCI) technology is to enable automatic emotion recognition capability within a multimodal dialog system framework. Most currently deployed spoken dialog interfaces, such as for instance, the ones used in call centers and intelligent tutoring, predominantly focus on the linguistic content, than user attitudes and emotions. These systems are limited in terms of handling the rich information contained in speech, and hence in their scope in terms of supporting natural human-machine interaction. Being able to detect the user's emotion can help us to enhance the capability of such interfaces in terms of being more natural and responsive to the user.

Most research on emotion recognition is primarily targeted to adult users even though children are one of the potential beneficiaries of computers with spoken interfaces, e.g., for educational applications and games. Recognizing a child's emotion during an interaction may help us to build interfaces that are better 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 and efficiency of the interaction. However, it is well acknowledged that the greater

[☆]Part of this work is presented at Interspeech-ICSLP 2005.

[*]Corresponding Author. Tel.: +90 326 6135600

Email addresses: serdar@mku.edu.tr (Serdar Yildirim), shri@sipi.usc.edu (Shrikanth Narayanan), potam@telecom.tuc.gr (Alexandros Potamianos)

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 (Narayanan and Potamianos, 2002). Automatic emotion recognition from speech is also a challenging research problem in many respects including, importantly, discerning the most appropriate signal features and classification methods. Given that most speech and linguistic characteristics of children vary with age and gender, it is important to identify emotionally salient features by means of emotion recognition as a function of gender and age group. In this paper, emotionally salient speech acoustic features and language usage as a function of age and gender are investigated.

There are databases of children speech that are mostly used for acoustic analysis and modeling. The KIDS corpus (Eskarnazi, 1996), American English CID children corpus (Lee et al., 1999), the CU Kids Audio Speech Corpus (Hagen et al., 2003) and the PF-STAR (British English, Italian, German and Swedish) (Batliner et al., 2005) can be listed as read speech corpora. Recently, databases of child-machine spontaneous speech interaction has been collected. The NICE database consists of open-ended spoken dialogue interaction between children and animated characters in a game setting (Bell et al., 2005). The FAU-AIBO corpus contains data from children spontaneously communicating with the AIBO robot (Batliner et al., 2006). Emotional labeling of FAU-AIBO corpus is also available to community. In this study, we used Children's Interactive Multimedia Project (ChIMP) database (Narayanan and Potamianos, 2002). The ChIMP database is a corpus of child-machine spoken dialog interaction in a game setting.

Acoustic features of speech have been used extensively to separate emotional coloring present in the speech signal by employing several pattern recognition techniques (Ang et al., 2002; Nwe et al., 2003; Lee and Narayanan, 2005; Batliner et al., 2006; Schuller et al., 2007b; Kapoor et al., 2007; Morrison et al., 2007; Neiberg and Elenius, 2008; Lee et al., 2009). Phoneme, syllable and word level statistics corresponding to F0 (fundamental frequency), energy, duration, spectral parameters, and voice quality parameters are among the features that have been mainly used for emotion recognition. Cowie et al. (Cowie et al., 2001) summarized acoustic correlates (in particular pitch, durations and intensity)for the archetypal emotions drawn from the vast body of literature on emotion. Recently, in Schuller et al. (Schuller et al., 2007a), a large number of acoustic features were grouped into Low Level Descriptor Types (LLD) and functionals and their impact on emotion classification performance has been analyzed in a comprehensive way.

Much of the work on emotion analysis and recognition focuses on databases with acted speech. Although research on acted emotional speech provides certain useful knowledge on how emotions are encoded in the speech signal, it is also important to work on data that are directly representative, and suitable, for the domain and application in mind. Furthermore, in some cases, it may be difficult to elicit acted speech. For instance, it is difficult to coach young children to produce acted speech. However, for many real world applications, including spoken dialog systems, it is not necessary to recognize a large set of emotions. As a consequence, the data coverage issues for seeking realistic, natural data also becomes somewhat more manageable. Thus, research on emotional data obtained from real applications has mostly focused on descriptions in a reduced emotion space such as negative vs. non-negative emotions (Lee and Narayanan, 2005), or frustration/annoyance vs. neutral (Ang et al., 2002) emotions. In this paper, we particularly focus on recognizing two attitudinal states of children in natural spontaneous spoken interactions, namely, *polite* and *frustrated* state, a problem which we believe is well suited to our application domain of child-computer interfaces. A preliminary study of politeness and frustration language in child-machine interactions was reported for different age groups in (Arunachalam et al., 2001). 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 politeness 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 *polite* and *frustrated* states in child-computer interaction.

In addition to acoustic features, lexical information can also be extracted from speech and used for emotion recognition, for example see (Ang et al., 2002; Lee and Narayanan, 2005; Seppi et al., 2008; Schuller et al., 2009a). Lee and Narayanan (Lee and Narayanan, 2005) proposed the notion of emotional salience, i.e., mutual information between a specific word and an emotion class, to select words in a speech utterance that are relevant for negative emotion detection. By augmenting the acoustic features with lexical information, a relative classification performance improvement of 46% was achieved. Similarly, in (Litman and Forbes-Riley, 2004), it was shown that the use of speech and language features for predicting student emotions in human-computer tutoring dialogs improved the accuracy of the system. Likewise, Zhang et al. (Zhang et al., 2006) reported promising results in the combined use of acoustic, spectral, and language information for detecting confidence, puzzlement, and hesitation in their child-machine dialog task. In Ang et al. (Ang et al., 2002), language model features measured from class-based trigram model were added

to prosodic decision trees; the results indicated that their language model features were poor predictors of frustration. In (Schuller et al., 2009a) vector space modeling and string kernels techniques were investigated for the recognition of emotion from spoken text. In our work, we extend the notion of emotional salience in language by calculating the mutual information between word pairs (bi-grams) and emotion classes.

In addition to mutual information based lexical feature extraction, we also propose to use Latent Semantic Analysis (LSA) (Deerwester et al., 1990) based feature extraction to obtain lexical information for emotion recognition. LSA is a technique based on Singular Value Decomposition (SVD) to construct a semantic space in which closely associated terms and documents are clustered together (Deerwester et al., 1990; Landauer et al., 1998). Chu-Caroll and Carpenter have successfully applied LSA to the problem of call-routing (Chu-Carroll and Carpenter, 1999). Our hypothesis here is that user utterances that express the same emotions will be associated with each other in the semantic space.

Discourse related information can also be used to predict user emotions. Several researchers have attempted to include discourse related information to improve emotion classification (Ang et al., 2002; Lee and Narayanan, 2005; Liscombe et al., 2005; Callejas and Lopez-Cozar, 2008). In (Ang et al., 2002), user turns were associated with a class of *repetition, correction*, or *else* and used as discourse features. Lee and Narayanan (Lee and Narayanan, 2005) used more discourse categories, *rejection, repetition, rephrase, ask-start over,* and *none-of the above*, to improve their negative/non-negative emotion detection task. In (Liscombe et al., 2005), in addition to discourse, contextual features related to changes in prosodic and lexical features between the current and previous user turn were also employed, with improved classification results. Callejas et al. investigated an influence of contextual information on emotion recognition performance for spoken dialogue systems (Callejas and Lopez-Cozar, 2008). Their results show that emotion recognition performance can be improved by using contextual information in addition to acoustic features. In the present study, dialog state and contextual information are used in conjunction with a variety of acoustic and lexical features.

The rest of this paper is organized as follows. Section 2 describes the database used in this study. Results from emotional state analysis are presented in Section 3. Section 4.1, 4.2, and 4.3 discuss acoustic, language, and contextual information sources, respectively. The information fusion algorithm is outlined in Section 5. Results are provided in Section 6 and Section 7 concludes the paper.

2. Database

The corpus used for analysis and modeling purposes in this paper is the Children's Interactive Multimedia Project (ChIMP) database (Narayanan and Potamianos, 2002). The ChIMP database is a corpus of spontaneous child-machine spoken dialog interaction in a game setting. The database contains speech data collected from 160 boys and girls, six to fourteen years of age. A Wizard of Oz (WoZ) technique was used for data collection that resulted in a database containing over 50,000 utterances. The task was to play "Where in the USA is Carmen Sandiego?", an interactive computer game using speech input. The goal of the game was to identify and arrest a cartoon criminal. During the game, the child had to interact with several animated characters to obtain clues about the suspect. Most children played the game twice. Games were labeled as successful when the child ended up arresting the suspect. Upon game completion children were asked to participate in an exit interview that gauged the interest of the child for the game and the speech interface. The collected acoustic data were transcribed and annotated with semantic and pragmatic information. Specifically, each user utterance was categorized into one of fifteen "dialog states" based on the requested action and game context. Further details about the database can be found in (Narayanan and Potamianos, 2002). Detailed linguistic analysis (duration, lexical and linguistic properties) of the database for each gender and age group is given in (Farantouri et al., 2008).

For the purposes of this work, we have annotated a subset of the ChIMP database, namely, data from 103 subjects out of the 160 total subjects with emotional state information. Specifically, each user utterance was labeled with one of three emotional state tags: *neutral*, *polite*, and *frustrated*. Each utterance was labeled independently by two native speakers of English. Labelers only took audio information into consideration. The agreement between the two annotators in terms of the Kappa statistic was 0.63. In this study, we consider 15585 utterances that both annotators agreed on. Also only data for children ages 7-14 are considered here (there was a very limited amount of data for age 6). Results are presented as a function of gender and three ages groups: 7-9 y/o, 10-11 y/o, and 12-14 y/o. The distribution of subjects for each gender and age groups is given in Table 1.

	Female	Male	Total
7-9 y/o	19	19	38
10-11 y/o	21	14	35
12-14 y/o	8	22	30
Total	48	55	103

Table 1: Number of subjects for each gender and age group.

	Neutral	Polite	Frustrated	Total
7-9 y/o	3966	977	796	5739
10-11 y/o	4004	1078	360	5442
12-14 y/o	3005	694	705	4404
Male	5940	1236	1061	8237
Female	5035	1513	800	7348
Total	10975	2749	1861	15585

Table 2: Number of instances (speaker turn) for each emotional class for each gender and age group.

3. Emotional State Analysis: Age and Gender Trends

In this section, the emotional state of the child is analyzed while interacting with the "Carmen Sandiego" computer game using a speech interface. The goal of the analysis of the ChIMP database is two-fold: (i) identify age and gender trends in emotional state, and (ii) identify lexical, semantic and pragmatic markers of emotional state. Preliminary results of this analysis can be found in (Arunachalam et al., 2001). The distribution of the emotional categories with respect to age group and gender are given in Table 2.

The first column of Table 2 provides the distribution of utterances per age and gender. Columns two to four show the number of utterances that were labeled as neutral, polite or frustrated for each age group and gender (only utterances where both labelers agree are shown). Note the significant age and gender trend in the statistics. As far as politeness is concerned, the 7-9 age group already shows high-levels[1] of polite behavior. This is consistent with research results from language acquisition showing that even six and seven year-old children have awareness and command of varying levels of politeness (Andersen et al., 1999). It has also been shown that children use impoliteness (insult) more frequently than adults when interacting with spoken dialog system (Bell, 2003). Children ages 10-11 are significantly more (often) polite than the other two age groups for this specific task. This trend is reversed for frustration. The younger (7-9) and older (12-14) age group appear about twice as frustrated as the middle (10-11) age group in their expressions. As far as gender is concerned, girls are significantly in expressing politeness and less often frustrated than boys, in their interactions with the computer characters.

To better understand the age and gender trends one should take into account additional factors, such as, game challenge, task completion and speech interface errors. Based on our analysis of interaction patterns it was clear that older children tended to complete the game faster, did fewer database lookups, used more advanced dialog patterns, and had fewer out-of-domain utterances than younger children. These observations together with the results from the exit interviews indicate that the game is challenging for the younger group, a good match for the skills of the middle age group and not much of a challenge for the older children. The frustration age trend can be partially attributed to the game challenge factor, i.e., children get frustrated when the game is too challenging or when the game is too easy for them. Task completion is also a factor; verbal expressions of frustration occurred more than twice as often in games that ended up in a loss than in those that were won (Arunachalam et al., 2001). A final comment has to do with speech recognition errors. In the data analyzed here, perfect speech recognition performance is assumed (using a wizard). From anecdotal results in a pilot study where recognition errors were randomly inserted, it is clear that frustration goes up significantly where recognition errors are involved (similar results are reported for adults (Gustafson and

[1]The fact that on average 15-20% of the utterances are classified as polite has also to do with the nature of the game that involves interaction with animated characters.

Bell, 2000)). Recognition errors seem to be irritating certain children much more than others. More details about emotional state age and gender trends, as well as, linguistic markers of emotional state are given next.

3.1. Lexical, Semantic and Pragmatic Markers of Politeness and Frustration

Explicit and implicit politeness markers were analyzed in the ChIMP data. Explicit markers included "please", "thank you", "excuse me", while implicit markers investigated were modals such as "may I", "could you", "would you". Younger children used simpler politeness constructs such as "excuse me", while they were not yet able to use implicit markers. Children ages 10-11 use overt politeness markers but do not yet fully employ polite request forms. Children ages 12-14 express politeness by a mix of explicit and implicit markers. It is also clear that the variability of politeness markers increases with age. We can speculate that the age trend in the level of politeness has to do with the nature of "social standing" that the child attributes to the animated characters, as well as, the challenge that the game provides. Two potential trends can be identified: (i) for younger children, the game is challenging and thus spend less time being cordial, and (ii) younger children hold the animated characters at a higher social standing thus being overtly polite. These two opposite trends can partially explain why the middle age group ends up being the most polite.

Next, we analyzed utterances labeled as "frustrated" to identify relevant lexical markers. Common verbal expressions of frustration, annoyance, and rejection were identified across gender and age group. Typical frustration markers included: "shut up", "oh man", "hurry", "oops", "heck". The usage of frustration markers varied with age and gender, as well as, with each individual child. Overall, children used less profanity than adults, but expressed frustration more often. In addition to lexical markers, pragmatic information served as a good feature for detecting frustration. Specifically, repetition or getting stuck in the same dialog state for multiple turns often indicated that a child was experiencing difficulty with the task and was getting frustrated.

Overall, girls are more polite and are less often frustrated than boys in spoken child-computer interaction. Some common "warning words" are especially salient in indicating "emotional" behavior. In addition, child age and gender significantly affects children's choices about politeness and frustration.

4. Feature Extraction and Modeling of Emotional State

In this section, feature extraction and modeling of the child's emotional state for spoken child-computer interaction is proposed. Motivated by the analysis of the child-machine interaction dialogs from the ChIMP corpus discussed in the previous section, acoustic, linguistic and dialog-based features are proposed, as well as, statistical models for characterizing the three emotional states of interest: neutral, polite and frustrated.

4.1. Acoustic Feature Extraction

We used the same low-level descriptors (LLD) and statistics that were proposed in (Schuller et al., 2009b). 384 features were extracted using openSMILE (Eyben et al., 2009) feature extraction. These features comprise of utterance level statistics corresponding to pitch frequency, root mean square (RMS) energy, zero-crossing-rate (ZCR) from the time signal, harmonics-to-noise ratio (HNR) by autocorrelation function, and 1-12 Mel-frequency cepstral coefficients (MFCC). Delta coefficients were also computed to each of these LLD. 12 statistics mean, standard deviation, skewness, kurtosis, maximum and minimum value, relative position, range and two linear regression coefficients with their mean square error were computed from each of these LLD and delta coefficients.

4.2. Lexical Feature Extraction and Modeling

As discussed in Section 3.1, certain words are associated with specific emotions and attitudes. Thus, lexical cues can be used to predict the emotional/attitudinal state of the user. Two different modeling approaches are proposed here to create lexical models of emotional state. First, an information-theoretic analysis is used for lexical feature selection (Cover and Thomas, 1991; Gorin, 1995; Lee and Narayanan, 2005), in conjunction with Bayesian classifiers for modeling. Second, latent semantic analysis (LSA) (Deerwester et al., 1990) is used to transform the feature space and then cosine distance metrics are used to compute "emotional distance" between utterances. Note that both information-theoretic analysis and latent semantic analysis are widely used techniques in text processing, indexing, and retrieval (Manning et al., 2008).

Male	Female	7-9y/o	10-11y/o	12-14y/o	Class
drop it	Hey you	do it	no thank	find the	Frustrated
get me	you there	stop miss	not that	pick that	Frustrated
shut up	someone talk	need this	my pad	go talk	Frustrated
stop this	you repeat	I don't	you pick	to issue	Frustrated
stop please	you mind	hello there	doing mister	hello I'd	Polite
you good	suspect can	please show	you have	thanks can	Polite
please tell	person please	very much	please take	would you	Polite
the phone	you can	you get	look that	where'd she	Polite

Table 3: Some salient word phrases and related emotion category for each gender and age group.

4.2.1. Lexical Feature Selection Using Emotional Salience

The notion of emotional salience was proposed in (Lee and Narayanan, 2005) for detecting negative and neutral emotions on real call center data. Lexical feature salience for emotion classification builds upon information-theoretic principles and feature selection algorithms. In essence, the relation between lexical items and emotion categories can be modeled by means of mutual information. The end goal here is to identify the most salient lexical items for each emotion category. Next we extend the work in (Lee and Narayanan, 2005) to include not only words, but also word pairs and phrases as potential lexical features.

Let f denote the lexical feature and $E = \{e_1, e_2, ..., e_k\}$ denote the emotional space, i.e., the set of emotional categories. Emotional salience of f in relation to E is defined as:

$$S(f) = \sum_{j=1}^{k} P(e_j|f) \log \frac{P(e_j|f)}{P(e_j)} \tag{1}$$

Note that emotional salience is the Kullback-Leibler distance between the posterior probability $P(e_j|f)$ of a class given feature f and the *a priori* probability of that class $P(e_j)$. In essence, emotional salience measures the distance between our knowledge before and after feature f was observed. Large distances indicate that f is very informative to the classification process and thus this feature should be selected.

After calculating the salience values of all lexical features, a salient word pair dictionary was constructed by only retaining word pairs that have greater salience values than a pre-chosen threshold (Lee and Narayanan, 2005) optimized on held-out data. In Table 3, salient word pairs f for each age group and gender are shown, along with their corresponding emotion category e_j (last column) for which the posterior probability $P(e_j|f)$ was maximized. It is interesting to note that the age and gender trends for the frustrated and polite lexical markers is consistent with the observations of Section 3.1, e.g., modals appear as a politeness marker only for the older children group.

Once the salient features were identified, a Bayesian classifier was built to determine the most probable emotion class for each utterance. Let $F = \{f_1, f_2, ..., f_l\}$ be the lexical features extracted from an utterance. We assume feature independence and use the Bayes classifier to maximize the posterior probability of an emotional class given the extracted features as follows

$$\hat{e} = \arg\max_j P(e_j|f_1, f_2, ..., f_l) = \arg\max_j \prod_{m=1}^{l} P(f_m|e_j)P(e_j) \tag{2}$$

where $\hat{e}$ is the emotion class that utterance with features F gets classified to. Note that if the logarithm of probability is used the formulation is equivalent to the one proposed in (Lee and Narayanan, 2005). The *a priori* $P(e_j)$ and observation probabilities $P(f_m|e_j)$ are computed during training using maximum likelihood estimation.

4.2.2. Latent Semantic Analysis

In this section, we propose Latent Semantic Analysis (LSA) as a method for modeling lexical information in the context of emotion classification. The first step is to construct a $n \times m$ term-document[2] matrix, $\mathbf{A}$, with elements a_{ij}

[2]Here the "document" is actually an utterance and "terms" are words.

Table 4: Sample dialog with the corresponding dialog state tags. Just the user portion of the sub-dialog is given.

User	Dialog State
Can I talk to him please?	Talk2Him
Tell me about the suspect	TellmeAbout
Can I see my choices for height?	EnterFeature
Tall for height	EnterFeature
Thank you	CloseBook
Can I talk to him please?	Talk2Him
Tell me where did the suspect go	WhereDid

representing the number of occurrences of word i in utterance j. Let $\mathbf{d} = \{d_1, d_2, ..., d_m\}$ denote the utterances and $\mathbf{w} = \{w_1, w_2, ..., w_n\}$ denote the words where m is the total number of the utterances and n is the total number of the unique words in the data. The matrix $\mathbf{A}$ is normalized so that each row (term) vector is of unit length. The latent-semantic space is obtained by employing SVD and only retaining the first l largest eigenvalues[3]. The matrix $\mathbf{A}$ can be decomposed using SVD as follows,

$$\mathbf{A} = \mathbf{U_0 S_0 V_0^T} \tag{3}$$

where $\mathbf{S_0}$ is the diagonal matrix that contains eigenvalues, and $\mathbf{U_0}$ and $\mathbf{V_0}$ are orthonormal matrices. After keeping the largest l eigenvalues, we obtain

$$\tilde{\mathbf{A}} = \mathbf{USV^T} \tag{4}$$

A pseudo-document vector, $\mathbf{d}$, is computed for each test utterance. $\mathbf{d}$ is a representation of an user utterance in the constructed latent semantic space. We first construct a vector $\mathbf{x}$, where each element x_i in the vector is the number of times the ith term occurred in the given test utterance. Then $\mathbf{d}$ can be calculated by means of the following operation (Deerwester et al., 1990):

$$\mathbf{d} = \mathbf{x^T U S^{-1}} \tag{5}$$

The cosine measure is used to determine the similarity between two utterances in the semantic space. Given two pseudo-document vectors, $\mathbf{d_1}$ and $\mathbf{d_2}$, the cosine distance between those two vectors is defined as:

$$\cos(\mathbf{d_1}, \mathbf{d_2}) = \mathbf{d_1} \cdot \mathbf{d_2} / \sqrt{||\mathbf{d_1}|| \, ||\mathbf{d_2}||} \tag{6}$$

where $\cdot$ signifies the dot product between two vectors.

To classify each test utterance into an emotion class we calculate the cosine similarity between this utterance and each of the training utterances. The test utterance is assigned to the emotion class of the most similar training utterance. When multiple such training utterances exist, we use majority voting to decide.

4.3. Discourse and contextual information modeling

As discussed in Section 3.1, discourse structure can also be a good marker of emotion, e.g., repetition may imply frustration. In this section, we use the manually transcribed dialog states of the ChIMP database (see Section 2) as information for the discourse structure of the interaction. A sample dialog fragment with the corresponding dialog state tags is given in Table 4. Note that here a total of nine dialog (super-) states are used. Detailed description and analysis of the dialog states can be found in (Potamianos and Narayanan, 1998).

In order to model the relationship between emotional state $e(t)$ and dialog state $d(t)$ a simple Bayesian model is used that assumes that the emotional state depends directly on dialog state history. Specifically, we assume that the emotional state of the current turn is directly dependent on the dialog state of current and the three previous turns.

[3]The value of l is an empirical question. We calculate the values of l for each gender and age group by maximizing classification performance on held-out data.

The selection of dialog state history $N = 4$ to model emotion was motivated by the fact that the perplexity of N-gram dialog models reaches a plateau for $N >= 4$ for the ChIMP data[4] (Narayanan and Potamianos, 2002).

Assuming independence among emotional states, an utterance $u(t)$ is classified into emotion class $\hat{e}(t)$ based on the following equation:

$$\begin{aligned} \hat{e}(t) &= \arg\max_j P(e_j|d(t), d(t-1), d(t-2), d(t-3)) \\ &= \arg\max_j P(d(t), d(t-1), d(t-2), d(t-3)|e_j)\ P(e_j) \end{aligned} \tag{7}$$

where $d(t)$ represents the dialog state for dialog turn t, e_j is the emotional state class label and j is the class index.

4.3.1. Contextual Information

Up to this point we have assumed independence between emotional states. Although this assumption simplifies our models and, especially, model parameter estimation, it is not always accurate. The emotional state of consecutive dialog turns is correlated, because emotions are persistent. In order to take this effect into account without changing the basic assumptions in our classifiers, we have employed a trick similar to the one used in speech recognition systems, namely use the derivative of the acoustic features as an extra feature.

Let us define the acoustic feature vector $a(t)$ of utterance $u(t)$ corresponding to dialog turn t; $a(t)$ consists of 94 different features as discussed in Section 4.1. Then the first difference between the current and two previous frames is used as additional features, namely, $b_1(t) = a(t) - a(t-1)$ and $b_2(t) = a(t) - a(t-2)$. Assuming independence between the dialog state and acoustic features, the acoustic context features were combined with the dialog state features in a joint classifier, as follows:

$$\begin{aligned} \hat{e}(t) &= \arg\max_j P(e_j|d(t), d(t-1), d(t-2), d(t-3), b_1(t), b_2(t)) \\ &= \arg\max_j P(d(t), d(t-1), d(t-2), d(t-3)|e_j)\ P(b_1(t), b_2(t)|e_j)\ P(e_j) \end{aligned} \tag{8}$$

where $\hat{e}(t)$ is the emotion class attributed to dialog turn t.

5. Fusion of Acoustic, Lexical and Contextual Information

In this work, we used decision level fusion to combine different information streams, namely, the acoustic, lexical and contextual information sources. Assuming that all single-stream classifiers are statistical and compute posterior probabilities, a popular fusion algorithm is to estimate the posterior probability of the combined classifier as a function of the posterior probabilities of the single-stream classifiers. Two common such fusion functions are the average and the product (assuming independence between streams).

Let $P(e_j|x_i)$ be the posterior probability estimate from each classifier, where e_j is the emotional state class label, j is the class index, x_i is the feature set from an information source i, and $i = 1, ..., N$ denotes the various information sources (acoustic, salient lexical features, LSA-based lexical features, dialog state and acoustic context features). Then the average fusion rule can be formulated as follows:

$$P(e_j|\mathbf{x}) = \frac{1}{N} \sum_i P(e_j|x_i) \tag{9}$$

i.e., the posterior probability of the combined classifier is the average posterior probability of the single stream classifiers (Chair and Varshney, 1986).

In our case, only some of the single stream classifiers are statistical and produce posterior probabilities, namely, salient lexical and context information classifiers, as can be seen in Eqs. (2) and (8).

[4] Although the selection of $N = 4$ is appropriate for the ChIMP child-computer interaction database where sub-dialogs rarely exceed four turns, shorter or longer dialog history and dialog state encodings might be more appropriate for other applications.

Feature	Gender		Age		
	Male	Female	7-9y/o	10-11y/o	12-14y/o
Lex1	478	478	425	399	353
Lex2	1502	1592	1265	1232	1036
LSA	240	200	240	240	260

Table 5: The total number of salient unigrams (Lex1), salient bigrams (Lex2), and the dimension of the "semantic" space (LSA) are shown for each gender and age group.

For the acoustic and LSA-based lexical classifier a distance metric is used instead for the classification decision. To transform these metrics into "posterior probabilities", the sigmoidal transformation of the (negated) distance metric was used. Specifically, we have used a modified version of the logistic function, as follows:

$$p(d) = \frac{1}{1 + c_1 e^{c_2\, d}} \tag{10}$$

where $p(d)$ is the pseudo-probability estimate for distance d, and c_1, c_2 are positive constants estimated on held-out data to maximize classification performance.

6. Experimental Results

In this section, we present experimental results on automatically detecting frustration, politeness and neutral attitudes in children's speech using the proposed features. We have identified the following classification tasks: acoustic feature evaluation, two-way classification (polite vs. others and frustrated vs. others), and three-way classification between all three emotion classes. Two-way classification is relevant for two reasons: (i) politeness represents an attitude in speaking style rather that an emotional state, while frustration reflects the user's emotional state, and (ii) there are applications where it is adequate to differentiate between polite vs. others or frustrated vs. others.

To investigate gender and age trends in classification, each age group and gender data were considered separately. The performance of the classifiers was evaluated by leave one speaker out (103-fold) cross-validation. We used three metrics to evaluate our classifiers. As classes are unbalanced in the database, we utilize unweighted average (UA) recall to evaluate classification performances. Unweighted average recall is the arithmetic mean of recall values of each emotion class. Unweighted average recall is reported for each experiment and overall for all gender and age groups. We also reported recall r and precision p values for each emotional classes. Recall r is defined as the ratio of the number of correctly classified instances over the total number of instances in the database for a class, and precision p is defined as the ratio of the number of correctly classified instances over the total number of classified instances for a class.

The following systems were evaluated:

- Support Vector Machine (SVM) classifiers were trained for emotion classification using Acoustic features (Acou), lexical features (single word) (Lex1) and bigram lexical features (Lex2). The training algorithm is implemented using libSVM (Chang and Lin, 2001) with linear kernels.
- Cosine-distance metric based classifier using latent semantic analysis and lexical features (LSA).
- Context based naive Bayes classifier using dialog state and acoustic context features (C).
- Synthetic Minority Oversampling TEchnique (SMOTE) (Chawla et al., 2002) was applied to obtain a more balanced class distribution for the training set and therefore to avoid classifier over-fitting. In particular, the number of instances in the polite (for 7-9 y/o and 12-14 y/o) and frustrated (for all age groups) classes were doubled for the three-way classification.
- Various two and three way combinations of the above systems using the posterior probability averaging for information fusion as discussed in Section 5.

The total number of salient unigrams used in Lex1, the total number of salient bigrams used in Lex2, and the dimension of the "semantic" space for the latent semantic analysis classifier (LSA) are given in Table 5, for each gender and age group. The cutoff point for Lex1, Lex2 and LSA features was estimated on held-out data so that three way classification performance was maximized.

	LLD				
	MFCC	F0	RMS Energy	Voicing	ZCR
Male	67.9	45.4	42.9	40.3	41.1
Female	70.4	51.3	44.9	45.8	46.3
7-9 y/o	70.4	51.7	44.2	45.7	44.2
10-11 y/o	66.4	47.9	45.2	44.0	43.6
12-14 y/o	70.6	49.3	44.0	43.0	44.9

Table 6: Performance comparisons of different acoustic feature groups for each gender and age group in terms of unweighted average recall (%).

6.1. Acoustic Features Evaluation

In order to evaluate the classification performance of each LLD, we have used a k-nearest neighborhood classifier (k-NNR) with $k = 3$. Classification results for the three categories (neutral, polite, frustrated) are computed using ten-fold cross validation. The motivation behind this analysis is to investigate the classification power of each low-level descriptors for emotion recognition as a function of age and gender. As can be seen from the Table 6, MFCC is the most discriminative low-level descriptor followed by F0. Results are consistent across gender and age groups.

Feature	Gender		Age		
	Male	Female	7-9y/o	10-11y/o	12-14y/o
Acou	75.0	79.5	75.2	80.9	75.9
Lex1	76.0	83.1	82.9	83.6	76.1
Lex2	75.4	81.9	78.7	82.4	75.3
LSA	77.5	82.6	83.4	84.1	77.6
C	62.9	70.1	68.5	68.5	62.2
Acou+Lex1	79.1	83.4	80.9	85.7	80.4
Acou+Lex2	77.8	82.2	79.4	83.7	80.1
Acou+ LSA	82.1	85.2	82.9	86.6	83.1
Acou+C	72.3	78.9	77.5	78.6	77.8
Acou+Lex1+C	80.1	84.7	82.4	87.8	83.8
Acou+Lex2+C	78.8	84.0	81.0	86.5	83.3
Acou+LSA+C	84.0	86.4	84.6	88.8	85.7

Table 7: Polite (P) vs. Others (O) classification results in terms of unweighted average recall (%) as a function of gender and age for acoustic (Acou), lexical (Lex1, Lex2, LSA), context (C) classifiers, and their fusion.

6.2. Two-way classification experiments

In the section, we present results for the polite vs. others and frustrated vs. others two-way classification experiments. Results for the polite vs. others task are reported in Tables 7 and 8 for unweighted average recall, and precision p and recall r, respectively. Results are reported as a function of gender and age group for acoustic (Acou), lexical (Lex1, Lex2, LSA) and context (C) feature classifiers, as well as, their fusion. Our goal is to investigate the performance of the different feature sets for politeness detection and identify possible age and gender trends.

It can be seen from Tables 7 and 8 that Lex1 features[5](unigrams) and LSA based lexical features provide the best performance for politeness detection, followed by the acoustic (Acou). This could be attributed to the limited

[5]The worse performance of the Lex2 (bigrams) features could be attributed to the limited amount of training data and the lack of smoothing.

		Male		Female		7-9 y/o		10-11 y/o		12-14 y/o	
		O	*P*	*O*	*P*	*O*	*P*	*O*	*P*	*O*	*P*
Acou	*r*	91.8	58.2	90.4	68.5	89.0	62.5	92.1	69.7	92.1	59.7
	p	92.7	55.1	91.7	65.1	92.0	53.8	92.3	69.2	92.7	57.6
Lex1	*r*	94.2	57.9	95.5	70.8	95.6	70.1	94.5	72.8	94.9	57.3
	p	92.8	63.2	92.6	80.2	93.9	76.8	93.2	77.0	92.5	66.7
Lex2	*r*	94.4	56.5	94.1	69.6	96.2	61.2	96.3	68.5	95.6	55.0
	p	92.6	63.4	92.3	75.6	92.4	77.0	92.3	82.4	92.2	69.3
Lsa	*r*	95.1	60.0	94.7	70.6	94.4	72.5	95.3	73.0	95.2	60.0
	p	93.2	67.7	92.5	77.5	94.3	72.5	93.3	79.9	93.0	369.3
C	*r*	94.7	31.1	92.6	47.6	92.5	44.5	95.0	42.0	93.1	31.3
	p	88.8	50.4	87.2	62.7	89.0	54.9	86.5	68.4	88.3	45.1
Acou+Lex1+C	*r*	92.4	67.7	92.5	76.9	90.1	74.7	94.3	81.2	92.9	74.6
	p	94.3	60.7	93.9	72.9	94.6	60.7	95.2	78.3	95.3	65.7
Acou+Lex2+C	*r*	91.5	66.0	91.1	76.8	88.4	73.7	92.6	80.4	92.5	74.2
	p	94.0	57.3	93.8	69.3	94.2	56.6	94.9	73.4	95.2	64.0
Acou+Lsa+C	*r*	93.8	74.1	94.4	78.4	91.9	77.2	95.6	82.0	93.8	77.6
	p	95.4	67.5	94.4	78.3	95.2	66.3	95.4	82.8	95.9	69.3

Table 8: Polite (*P*) vs. Others (*O*) classification results in terms of precision (*p*) (%) and recall (*r*) (%) as a function of gender and age for acoustic (Acou), lexical (Lex1, Lex2, LSA), context (C) classifiers, and their fusion.

Feature	Gender		Age		
	Male	Female	7-9y/o	10-11y/o	12-14y/o
Acou	68.6	69.4	69.2	65.9	73.8
Lex1	62.4	58.5	59.9	50.4	63.7
Lex2	56.1	54.3	49.8	50.1	58.5
LSA	59.8	60.5	57.5	50.0	65.9
C	65.1	65.8	66.7	64.7	71.7
Acou+Lex1	69.2	69.3	69.8	65.1	73.5
Acou+Lex2	67.9	68.5	68.8	64.2	69.8
Acou+ LSA	68.0	68.4	69.1	62.3	73.1
Acou+C	70.0	70.3	71.7	67.8	75.3
Acou+Lex1+C	70.7	71.4	72.2	69.3	75.1
Acou+Lex2+C	70.3	71.6	71.9	69.1	75.5
Acou+LSA+C	70.9	71.4	71.9	69.5	75.6

Table 9: Frustrated (*F*) vs. Others (*O*) classification results in terms of unweighted average recall (%) as a function of gender and age for acoustic (Acou), lexical (Lex1, Lex2, LSA), context (C) classifiers, and their fusion.

lexical variability of politeness constructs in this database, i.e., politeness markers are limited to a few highly frequent phrases such as *please, thank (you), excuse (me)*. The two-way and three-way fusion of classifiers resulted in improved performance with best results being achieved with the combination of acoustic, lexical (unigrams) and contextual information (Acou+Lex1+C).

Overall, politeness detection is somewhat more accurate for female speakers and for the middle age group (ages 10-11) as can be seen in Table 7. This trend can be seen more clearly in Table 8 where the recall for the polite class is significantly higher for females and the middle age group, which implies that the politeness markers are more clearly identifiable for these gender and age groups.

For the frustrated vs. others classification task the results are shown in Tables 9 and 10 for unweighted average recall, and precision *p* and recall *r*, respectively. Note that the results are directly comparable to the politeness detection task, since the same classifier was used in both experiments. One can observe that for frustration detection the best results were obtained for the acoustic (Acou) and context (C) classifiers, followed by the lexical classifiers

		Male		Female		7-9 y/o		10-11 y/o		12-14 y/o	
		O	*F*	*O*	*F*	*O*	*F*	*O*	*F*	*O*	*F*
Acou	*r*	95.7	41.1	97.9	38.5	96.4	42.1	98.7	33.1	94.2	53.4
	p	91.6	59.0	92.8	69.3	91.1	65.5	95.4	65.0	91.3	64.2
Lex1	*r*	96.5	28.2	96.9	20.2	96.6	23.2	97.9	02.8	94.9	32.5
	p	90.4	53.7	90.8	44.6	88.6	52.3	93.4	08.9	87.9	55.3
Lex2	*r*	97.9	14.2	98.4	10.1	95.8	03.8	99.0	01.1	97.0	20.0
	p	88.4	49.8	89.9	43.9	85.9	12.9	93.3	07.8	86.2	56.9
Lsa	*r*	96.5	23.1	97.3	23.7	97.1	17.9	97.5	02.5	95.4	36.5
	p	89.4	49.9	91.2	52.2	87.9	49.8	93.3	06.8	88.6	60.5
C	*r*	96.3	33.9	97.7	33.9	96.6	36.9	98.8	30.6	93.8	49.6
	p	90.7	57.9	92.3	64.8	90.4	64.1	95.2	66.3	90.6	60.9
Acou+Lex1+C	*r*	96.1	45.4	98.0	44.9	96.6	47.8	98.7	39.9	94.6	55.5
	p	92.2	63.2	93.5	73.3	91.9	69.3	95.5	69.4	91.6	66.7
Acou+Lex2+C	*r*	96.2	44.5	98.1	45.1	96.6	47.3	98.7	39.4	94.8	56.3
	p	92.1	63.7	93.6	74.4	91.9	69.3	95.8	69.2	91.8	67.6
Acou+Lsa+C	*r*	95.9	45.9	98.1	44.9	96.5	47.4	98.8	40.2	94.5	56.7
	p	92.2	62.7	93.5	73.6	91.9	69.0	95.6	70.3	91.8	66.8

Table 10: Frustrated (*F*) vs. Others (*O*) classification results in terms of precision (*p*) (%) and recall (*r*) (%) as a function of gender and age for acoustic (Acou), lexical (Lex1, Lex2, LSA), context (C) classifiers, and their fusion.

Feature	Gender		Age		
	Male	Female	7-9y/o	10-11y/o	12-14y/o
Acou	61.4	63.7	62.2	60.5	61.9
Lex1	61.1	63.7	62.9	56.4	61.5
Lex2	55.6	60.3	54.5	55.1	60.2
LSA	60.3	63.9	61.9	56.6	63.8
C	49.3	54.1	54.9	51.6	51.5
Acou+Lex1	63.0	65.9	64.9	62.6	64.4
Acou+Lex2	62.0	65.6	63.7	62.1	64.1
Acou+ LSA	65.5	67.5	66.8	63.7	66.2
Acou+C	65.2	61.3	61.1	57.5	51.5
Acou+Lex1+C	63.8	67.3	66.1	63.5	66.4
Acou+Lex2+C	62.7	66.9	64.6	63.4	66.8
Acou+LSA+C	66.8	68.8	67.2	65.5	68.3

Table 11: Three-way classification results in terms of unweighted average recall (%) as a function of gender and age for acoustic (Acou), lexical (Lex1, Lex2, LSA), context (C) classifiers, and their fusion.

(Lex1, LSA, Lex2). Note that contextual information has more discriminative power than lexical information for all gender and age groups. This can be explained by the fact that frustration builds over multiple turns and is sometimes identifiable by discourse markers such as dialog state repetition. Also there is greater lexical overlap between neutral and frustrated speech than between neutral and polite speech utterances (see also (Ang et al., 2002)).

In the two-way fusion of classifiers, the best result was achieved with the combination of acoustic and contextual information (Acou+C). Also note that three-way fusion of classifiers provide small improvement over the fusion of acoustic (Acou) and context (C) classifiers. There is no clear gender trend for frustration detection, other than a slightly better overall classification performance for females shown in Table 9. Also, the classification results are better for the 12-14 age group in terms of unweighted average recall. Overall, the recall results for the frustrated class are low for all classifier setups.

		Male			Female		
		N	*P*	*F*	*N*	*P*	*F*
Acou	*r*	80.2	52.9	51.1	85.8	62.8	42.4
	p	84.2	65.9	34.4	84.6	75.3	34.5
Lex1	*r*	85.2	61.2	36.9	87.8	70.9	32.3
	p	83.4	63.6	40.1	84.5	78.9	34.0
Lex2	*r*	86.0	57.1	23.6	86.5	70.6	23.7
	p	80.9	63.9	30.6	82.6	76.2	28.3
Lsa	*r*	87.1	61.7	32.2	87.6	71.0	33.1
	p	82.9	66.8	39.2	84.8	75.3	36.6
C	*r*	90.1	27.7	30.0	92.8	46.3	23.3
	p	78.6	59.9	37.5	78.7	70.9	43.3
Acou+Lex1+C	*r*	81.2	53.8	56.4	87.6	64.9	49.3
	p	85.5	67.9	37.1	86.0	79.9	39.9
Acou+Lex2+C	*r*	80.8	51.9	55.4	81.9	56.0	56.0
	p	85.1	64.3	36.9	85.2	66.1	40.7
Acou+Lsa+C	*r*	82.0	61.3	56.9	88.3	67.7	50.3
	p	86.3	74.4	38.6	86.5	86.2	39.5

Table 12: Performance of classifier in terms of precision (p) (%) and recall (r) (%) for each class for each gender. *N*:Neutral, *P*:Polite, *F*:Frustrated.

		7-9 y/o			10-11 y/o			12-14 y/o		
		N	*P*	*F*	*N*	*P*	*F*	*N*	*P*	*F*
Acou	r	80.2	56.1	50.2	90.6	65.8	26.3	84.2	65.9	34.4
	p	83.9	63.2	37.0	86.8	76.1	26.3	73.6	54.5	57.8
Lex1	r	86.0	70.2	32.6	90.3	70.7	08.2	82.1	59.2	43.0
	p	82.3	76.9	36.9	86.4	77.0	10.9	80.3	65.3	43.2
Lex2	r	82.3	62.3	18.8	93.2	70.2	02.0	78.3	54.4	47.7
	p	78.8	73.5	19.4	84.9	79.7	06.5	80.3	65.6	37.8
Lsa	r	84.9	72.5	28.5	90.0	72.9	06.5	83.8	62.1	45.6
	p	82.2	69.2	36.4	86.6	79.6	07.9	81.5	69.1	46.6
C	r	91.2	42.8	30.7	93.9	40.4	20.4	87.5	27.0	39.2
	p	78.5	65.1	50.0	81.0	76.1	29.8	77.6	59.7	40.9
Acou+Lex1+C	r	83.1	58.4	56.9	94.6	68.5	27.5	81.5	51.5	66.2
	p	85.9	71.5	41.1	88.2	87.8	31.3	84.7	78.6	44.6
Acou+Lex2+C	r	81.9	56.0	56.0	93.9	65.8	30.3	81.3	52.1	66.9
	p	85.2	66.1	40.7	87.8	82.6	35.1	85.1	78.2	44.6
Acou+Lsa+C	r	84.3	61.6	55.6	96.4	70.2	30.0	81.5	57.1	66.2
	p	86.1	76.5	41.6	89.0	92.8	30.4	84.9	81.6	45.6

Table 13: Performance of classifier in terms of precision (p) (%) and recall (r) (%) for each emotion class for each age group. *N*:Neutral, *P*:Polite, *F*:Frustrated.

6.3. Three-way classification

In this section, the performance of the proposed features is investigated for the three-way classification problem as a function of gender and age group. Results are shown in Table 11 for unweighted average recall, and in Tables 12 and 13 for the recall and precision values for each emotion class. In terms of UA recall, the LSA-based features provide the best accuracy, closely followed by the acoustic (Acou) and unigram features (Lex1), while bigrams (Lex2) and context features (C) perform worse. Significantly better performance is reached when the acoustic and lexical features (LSA or Lex1) are fused (one-way ANOVA $p < 0.001$). By combining acoustic, lexical and context features a 5.4% absolute improvement over using only acoustic features (Acou) is achieved (Acou+LSA+C).

A gender and age trend can also be observed. It is clear from Table 11 that classification performance is higher for

females. As far as precision and recall are concerned in Table 12, there is gender dependency for the polite class, but not for the frustrated class. In Table 13, precision and recall values are poor for the middle age group for frustration even though Synthetic Minority Oversampling TEchnique (SMOTE) was applied to avoid classifier over-fitting.

7. Conclusions

An essential step toward building natural and responsive spoken interaction systems, especially for children, is to analyze and detect age- and gender-dependent user behavior patterns. In this paper, we analyzed the polite and frustrated behavior of children during spontaneous spoken dialog interaction with computer characters in a computer game. The analysis showed that girls and children aged 10-11 were significantly more polite and less frustrated than older and younger children, although the age trend could also be application dependent. The analysis also showed that some common "warning words" were especially salient in indicating polite and frustrated behavior. In addition to lexical markers, pragmatic markers, e.g., repetition, were often good indicators of frustration.

Next we investigated how acoustic, lexical and contextual information could be used for politeness detection, frustration detection and emotional state classification. To obtain lexical information in the context of emotion recognition, we proposed to use Latent Semantic Analysis (LSA) based feature extraction. The performance of LSA features was compared with that of word and word pair features selected via a mutual information, emotional salience criterion. Overall, single word features (Lex1) and LSA performed equally well for both detection tasks (two-way classification) and three-way classification. We also investigated contextual features that take into account dialog state information and turn-to-turn change of acoustic features. It was shown that the contextual features provide good performance especially for the frustration detection task.

Results show that lexical cues have more discriminative power than acoustic and dialog cues for detection of politeness, whereas dialog and acoustic cues are better for frustration detection. This is in agreement with the analysis results that show that politeness is more explicitly marked in language usage, while repetitions and corrections (due to system errors or task difficulty) may lead to frustration. Based on the results of both two-way and three-way classification experiments it is clear that by augmenting acoustic features with lexical and contextual information classification performance improves significantly. The results also showed age and gender trends, e.g., classification performance was better for girls than for boys.

There are several issues that must be further explored in the future. Pragmatic and task information such as the number of user turns and task (or subtask) success are some features that could be used to improve classification performance. A preliminary data analysis showed that there is a positive correlation ($r=0.259$, $p=0.001$) between the number of total dialog turns and child frustration. Similarly, task success affects level of frustration; there was a higher percentage of frustrated turns in unsuccessful games (0.7% vs 0.5%). Finally, there is negative correlation between the numbers of frustrated and polite user turns in an interaction (Kendall's tau-b=-2.87 $p<0.01$). Other sources of information that can be exploited for emotion classification include syntax, language register and disfluencies. Another important area of research is improved model and classifier design and classifier fusion, e.g., vector space representation for LSA, bigram smoothing, maximum entropy fusion. While this paper has taken some initial steps toward emotion detection and emotional state classification in natural spoken dialog child-computer interaction, numerous open research questions remain both in the analysis and modeling fronts.

References

Andersen, E., Brizuela, M., DuPuy, B., Gonnerman, L., 1999. Cross-linguistic evidence for the early acquistion of discourse markers as register variable. Journal of Pragmatics (31), 1339–1351.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A., 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: Proceedings of ICSLP. Denver, pp. 2037–2039.

Arunachalam, S., Gould, D., Andersen, E., Byrd, D., Narayanan, S., 2001. Politness and frustration language in child-computer interactions. In: Proceedings of Eurospeech. Aalborg, Denmark, pp. 2675–2678.

Batliner, A., Blomberg, M., D'Arcy, S., Elenius, D., Giuliani, D., Gerosa, M., Hacker, C., Russell, M., Steidl, S., Wong, M., 2005. Children's speech recognition with application to interactive book and tutors. In: Proceedings of InterSpeech. pp. 2761–2764.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V., 2006. Combining Efforts for Improving Automatic Classification of Emotional User States. In: Erjavec, T., Gros, J. Z. (Eds.), Language Technologies, IS-LTC 2006. Ljubljana, Slovenia, pp. 240–245.

Bell, L., 2003. Linguistic adaptations in spoken humancomputer dialogues - empirical studies of user behavior. Ph.D. thesis, KTH, Sweden.

Bell, L., Boye, J., Gustafson, J., Heldner, M., Lindstrom, A., Wiren, M., 2005. The Swedish NICE Corpus Spoken dialogues between children and embodied characters in a computer game scenario. In: Proceedings of InterSpeech. Lisbon, Portugal, pp. 2765–2768.

Callejas, Z., Lopez-Cozar, R., 2008. Influence of contextual information in emotion annotation for spoken dialogue systems. Speech Communication 50 (5), 416 – 433.

Chair, Z., Varshney, P., Jan. 1986. Optimal data fusion in multiple sensor detection systems AES-22 (1), 98–101.

Chang, C.-C., Lin, C.-J., 2001. LIBSVM: a library for support vector machines. Software available at `http://www.csie.ntu.edu.tw/ cjlin/libsvm`.

Chawla, N. V., Bowyer, K. W., Kegelmeyer, P. W., 2002. Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357.

Chu-Carroll, J., Carpenter, B., 1999. Vector-based natural language call routing. Computational Linguistics 25 (3), 361–388.

Cover, T., Thomas, J., 1991. Elements of Information Theory. John Wiley & Sons, New York, NY.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J., 2001. Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine 18 (1), 3280.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., Harshman, R. A., 1990. Indexing by latent semantic analysis. Journal of the American Society of Information Science 41 (6), 391–407.

Eskarnazi, M., 1996. Kids: A database of children's speech. Journal of the Acoustical Society of America 100 (4), 2759.

Eyben, F., Wllmer, M., Schuller, B., 2009. openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In: 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction (ACII).

Farantouri, V., Potamianos, A., Narayanan, S., 2008. Linguistic Analysis of Spontaneous Children Speech. In: Berkling, K., Giuliani, D., Potamianos, A. (Eds.), Proceedings of the 1st Workshop on Child, Computer and Interaction. Chania, Greece.

Gorin, A., 1995. On automated language acquisition. J. Acoust. Soc. Am. 97(6), 3441–3461.

Gustafson, J., Bell, L., 2000. Speech technology on trial - experiences from the august system. Natural Language Engineering 6 (3-4), 273–286.

Hagen, A., Pellom, B., Cole, R., 2003. Children's speech recognition with application to interactive book and tutors. In: Proceedings of the ASRU Workshop. pp. 186–191.

Kapoor, A., Burleson, W., Picard, R. W., 2007. Automatic prediction of frustration. International Journal of Human-Computer Studies 65, 724–736.

Landauer, T. K., Foltz, P. W., Laham, D., 1998. Introduction to latent semantic analysis. Discourse Processes 25, 259–284.

Lee, C.-C., Mower, E., Busso, C., Lee, S., Narayanan, S., 2009. Emotion recognition using a hierarchical binary decision tree approach. In: Proceedings of InterSpeech. Brighton, UK, pp. 320–323.

Lee, C. M., Narayanan, S., 2005. Towards detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing 13 (2), 293–302.

Lee, S., Potamianos, A., Narayanan, S., 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. Journal of the Acoustical Society of America, 1455–1468.

Liscombe, J., Riccardi, G., Hakkani-Tur, D., 2005. Using context to improve emotion detection in spoken dialog systems. In: Proceedings of Interspeech. Lisbon, Portugal, pp. 1845–1848.

Litman, D. J., Forbes-Riley, K., 2004. Predicting student emotions in computer-human tutoring dialogues. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL). Barcelona, Spain, p. 351.

Manning, C., Raghavan, P., Schutze, H., 2008. Introduction to Information Retrieval. Cambridge University Press, Cambridge.

Morrison, D., Wang, R., Silva, L. C. D., 2007. Ensemble methods for spoken emotion recognition in call-centers. Speech Communication 49 (5), 98–112.

Narayanan, S., Potamianos, A., 2002. Creating conversational interface for children. IEEE Transactions on Speech and Audio Processing 10 (2), 65–78.

Neiberg, D., Elenius, K., 2008. Automatic recognition of anger in spontaneous speech. In: Proceedings of InterSpeech. ISCA, pp. 2755–2758.

Nwe, T. L., Foo, S. W., Silva, L. C. D., 2003. Speech emotion recognition using hidden markov models. Speech Communication 41 (4), 603–623.

Potamianos, A., Narayanan, S., 1998. Spoken dialogue systems for children. In: Proceedings of ICASSP. Seattle, WA, pp. 197–201.

Schuller, B., Batliner, A., Seppi, D., Steidl, S., 2009a. Emotion Recognition from Speech: Putting ASR in the Loop. In: ICASSP (Ed.), IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4585–4588.

Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., 2007a. The relevance of feature type for automatic classification of emotional user states: Low level descriptors and functionals. In: Proceedings of InterSpeech. pp. 2253–2256.

Schuller, B., Seppi, D., Batliner, A., Maier, A., Steidl, S., 2007b. Towards more reality in the recognition of emotional speech. In: Proceedings of ICASSP. Vol. 4. pp. 941–944.

Schuller, B., Steidl, S., Batliner, A., 2009b. The INTERSPEECH 2009 Emotion Challenge. In: ISCA (Ed.), Proceedings of InterSpeech. pp. 312–315.

Seppi, D., Gerosa, M., Schuller, B., Batliner, A., Steidl, S., 2008. Detecting Problems in Spoken Child-Computer-Interaction. In: Berkling, K., Giuliani, D., Potamianos, A. (Eds.), Proceedings of the 1st Workshop on Child, Computer and Interaction.

Zhang, T., Hasegawa-Johnson, M., Levinson, S. E., 2006. Cognitive state classification in a spoken tutorial dialogue system. Speech Communication 48, 616–632.