An Empathic-tutoring System Using Spoken Language

Tong Zhang, Mark Hasegawa-Johnson, Stephen E. Levinson
University of Illinois at Urbana-Champaign
{tzhang1, hasegawa, sel}@ifp.uiuc.edu

Abstract

This paper represents an approach to simulate the mental activities of children during their interaction with computers through their spoken language. The mental activities are categorized into three states: confidence, confusion and frustration. Four knowledge sources are used in the detection. One is prosody, which is indicative of utterance type and user's attitude. The second is embedded key words/phrases which help interpret the utterances. Moreover, it is found that children's speech exhibits very different acoustic characteristics from adults. Given the uniqueness of children's speech, this paper applies a vocal-tract-normalization (VTLN)-based technique to compensate for both inter-speaker variability and intra-speaker variability in children's speech. The detected key words/phrases are then integrated with prosodic information as the cues for the MAP decision of mental states. Tests on a set of 50 utterances collected from the project preliminary experiments showed the classification accuracy was %.

1. Introduction

1.1. Empathic-tutoring system

An empathic-tutoring system is devised for an NSF-funded project, titled "Multimodal Human Computer Interaction System: Toward a Proactive Computer." The testbed of this project is an intelligent-tutoring environment for education in science and technology, using the Lego construction set, with children of primary and middle school age. The emphasis is on developing a proactive computer agent to encourage the interests of kids in science and technology. The communication between the computer agent and the user is via a spoken dialogue system. The project is studied based on three assumptions:

- the proactive assumption: The computer can initiate communications when the user needs guidance, assistance, and encouragement rather than simply waiting for user commands.
- the human-centred assumption: Children take initiative to discover and learn by themselves, while the computer plays auxiliary functions.
- the empathic assumption: The computer can detect the cognition activities of students during their learning process.

The proactive assumption and human-centred assumption require the computer to keep track of the users closely, detecting their status including the utterance contents, task stages, attention foci, and so on. The computer obtains these kinds of information from multiple sensors: speech, eye tracking, facial expression, body gesture, and hand action. In addition, the capability of detecting students' cognitive activities is prominent for a successful human tutor. Since the goal of education is to leading students going further and farther in cognition of the physical world, making clear students' current cognitive status is the key issue for going the next step. The empathic assumption requires the computer have more of this "human sense". This paper focuses on the empathic aspect of the intelligent tutoring project, so we call it an "empathic tutoring system."

1.2. Tutoring content

At the start of the game, the computer asks users a few personal questions, such as "How old are you?" and "Have you played with Legos before?", to obtain a rough estimation of the user's knowledge base. The tutoring consists of basic mathematics, physics and mechanical engineering.

First, mathematics can be learned through manipulating concretes rather than solely handling abstract symbols (Wilensky, 1991). So we try a novel method of primary education in Mathematics—playing with Lego gears. In our system, the child users are given gears of different size. The teeth on each gear are paint with different colors: red and blue, red and green, or blue and green. The child users are asked to design the playing strategy themselves to finally answer several questions. Some questions are based on the mathematical calculations. For example, one question is based on the ratio of teeth number and spinning circles:

Line up a 24-tooth gear and a 40-tooth gear. If the 24-tooth gear spins 5 times, then how many times must the 40-tooth gear spin for them to line up again? and why?

Children are expect to line up a 24-tooth gear and a 40-tooth gear along a beam and right next to each other, and then swirls those gears, counting and comparing the spinning cycles. Kids find that gears with more teeth spin more slowly. Some kids can further figure out that the product of teeth number and spinning circles is the same for the two gears. The traditional approach of teaching basic mathematics is through memorizing formulas and rules about "inversely proportional." This system provides children concrete objects (Lego) so that users can develop the physical sense of this abstract definition. Or else, if they already know the abstract definition of "inversely proportional," they can connect it to this physical model to solidify the mathematical knowledge they have learned. It can also excite child's interests using mathematical knowledge to resolve practical problems.

Second, physics and mechanical engineering are disciplines involving more direct interaction with the physical world. So some basic knowledge of them can be learned by playing Lego also. For example, a question about "interactive force", a physics concept, is

Put one hand on the 40-tooth gear axel, and put another hand on the 8-tooth gear axle. What happens if you hold one of them steady, and try to turn the other one? And why?

The third implication of "tutoring" is that the computer agent initiates a self-tutoring process by educating child users. It is proposed that nonexperts can learn by consulting other lower-level nonexperts (Kafai and Harel, 1991). So a computer agent is expected to expand its knowledge base by helping student users in their learning.

1.3. Defining cognitive states

Motivation plays important roles in learning (Galbraith, 1998). Human cognition is a continuous process, but we can discretize the continuous cognition activity according to the stages of motivation. In this system, we use cognitive states to encode the motivation stages of child users during the process of knowledge acquisition by problem solving. According to the preliminary study of the project, the cognitive states of children generally can be summarized into three cases: confidence, confusion and frustration. Cognitive state detection is pragmatically useful for successful interaction: it provides clues for the response strategy of computer agent. When the user is confident about his/her action, e.g., "I see the small one moving faster," the computer agent needs to initiate communications by providing suggestion for his/her next action. When the user is confused, e.g., "What do you mean?" the computer agent needs to help him/her get a better understanding or provide guidance by asking open questions. When the user is frustrated or hesitant, e.g., "I count eight then for the … sort of one, but …" the computer agent needs to encourage or help the user clarify his/her ideas.

2. Background

The intelligent-tutoring project originates from Constructionism, "a theory of learning and a strategy for education" (Kafai and Resnick, 1996). Constructionism proposes that children can learn knowledge not solely by straightforward instruction. Devising some type of "external artifact" is another way of learning knowledge: practice provides an efficient way for children to think and motivate new ideas themselves. Aside from active learning theory, the AutoTutor system manifests that knowledge can be gained not only by instruction, but also by conversations (Graesser, 2001). In fact, AutoTutor simulates the conversation of people, but it is semi-conversational: the tutor exchanges information through a computer-generated head talking with synthesized speech, while the students respond by typing. Aiming to help college students master basic computer techniques, the experimental results showed that students can improve their learning by 0.5 standard deviation units comparing with learning by reading.

As the greatest intelligence indicator, spoken language contains rich information which can reflect the intra-speaker variation, caused by the speaker's brain activities, during communication. An outstanding indicator of the intra-speaker variation is emotion. This is because the emotional activity causes physiological variation in the vocal mechanism, which is used to generate sound, and further causing speech variation. For example, when people are in fear or anger, the muscular system would be in tension. The tension of muscles in the vocal mechanism raises the fundamental frequency of sound (Stevens, 1991). So the rising fundamental frequency (pitch) is a nonverbal indicator of fear and anger. Aside from prosody, affect bursts, "a very brief, discrete, nonverbal expressions of affect in both face and voice as triggered by clearly identifiable events," is shown to reveal emotional meanings (Schröder, 2003). Moreover, acoustic characteristics are proven to be related with mental activities by experimentally measuring the transient variation of electroencephalogram (EEG) voltages, an indicator of brain responses triggered by events (Alter, et al., 2003).

As the important aspects of speech, prosody and lexicon are widely used in the intra-speaker variation-related issues. The most prevalent issue is emotion. Great efforts are made in emotional issues via speech. For example, prosody and lexicon are used to study the

speaker's happiness, ?, and ? in their interaction with computers (Pozin and Waibel, 2001).

Another perspective looking through the issue of intra-speaker variation is behaviour. Prosody and lexicon are also related with and thereafter applied to detecting the behavioural variation of speakers. Studies reveal that the utterance duration tends to be longer in the speech of error corrections (Levow, 1998). The repetition speech tends to have higher magnitude and longer utterance duration (Bell and Gustafson, 1999). The increase of speech magnitude also occurs in repetition speech. And also, speakers tend to enlarge the phonetic word boundaries between the correction and its adjacency (Menezes, et al., 2003). Word information is used to detect colloquial speech repairs, speech repetition, intonational phrases and discourse markers (Heeman and Allen, 1999). The lengthening of the pre-boundary final syllable manifest intonational phrases (Edwards, 1991). In a human-computer interactive system, prosody and lexicon can be used for dialogue act classification, such as Yes-No question, statement, declaration, and so on (Jurafsky, et al., 1997). Word repetition and some prosodic features are used to find trouble in the call centre, so that the call is transferred to a human operator in time (Batliner, et al., 2003). In addition, prosody and lexicon are more recently used in the automatic meeting documentation, such as spoken sentence segmentation and topic segmentation (Shriberg, et al., 2000).

3. Prosodic analysis

Summarized into Table 1, the prosodic features used in the mental state discrimination are divided into several types:

- Pitch: F0 is derived using a typical autocorrelation method, but with the expected pitch range adjusted in order to account for the disparity between children's speech and adult speech (Lee, et al., 1999). Confused children tend to ask more questions; furthermore, confused children tend to exaggerate the pitch-rise at the end of their questions, while confident children tend to exaggerate the turn-final declaration fall. Unlike either confidence or confusion, in some cases the pitch contour at the end of frustrated sentences is neither rising nor falling.
- Voice-unvoiced percentage: The percentage of time that voiced speech occupies in an utterance is a good cue for discriminating frustration from confidence and confusion.
- Energy-related features: Typically, an utterance falls to lower energy when close to completion. Nevertheless, when speakers stop mid-stream, this fall has not yet occurred and thus energy remains unusually high (Jurafsky, et al., 1997). Therefore, the comparison of energy in the end and nearly-end regions can be indicative of frustration. The log-scaled energy normalized by the peak and its first and second order time derivates at the end of the utterance are also investigated.
- Pause-related features: As a cue of frustration, pause refers to a time period of non-speech signals over 600ms. If the non-speech duration is less than 600ms, it is more likely to be a natural transition within sentences. To detect a pause, we use the time-varying threshold method of Li et al. (2002).
- Syllabic rate-related features: Syllabic rate is defined as the number of syllables normalized by the utterance duration. Usually people speak more slowly when they are hesitant and confused, as opposed to when they are confident. The speaking rate is computed by the average of three estimators. The first estimator is peak counting performed on the wide-band energy envelope, and then normalized by the utterance duration. The second estimator is a sub-band-based module proposed by Morgan and Fossler-Lussier (1998). Our third estimator is the modulation-spectrum method of Kitazawa et al. (1997). Since the speaking rate of people is usually different, the syllabic rate of an utterance is normalized by the speaker's average syllabic rate.
- Word duration-related features: If the duration of a word is longer than the duration of other words, and at the same time the word is not accentuated, then the word is a possible frustration indicator and locator. The word duration is a product of speech recognition, which is recounted in the next section. Word duration can be studied from both absolute and normalized versions (Batliner, et al., 2002). In this study, the word duration considers simultaneously average and maximum of absolute duration and duration normalized by number of syllables, as well as utterance duration normalized by the number of syllables in the utterance.

Table 1. The prosodic features used in the mental state classification

Features	Description
F0_ratio	ratio of mean F0 over the end region (the final 100ms) and the penultimate region (the previous 100ms).
F0_reg_pen	least-square all-points regression over the penultimate region.
F0_reg_end	least-square all-points regression over the end region.
F0_norm	the number of frames with non-zero F0 in an utterance normalized by the utterance duration.
logE_ratio	ratio of logarithmic energy over the end region and the penultimate region
Derive_logE	mean of peak-normalized

	logarithmic energy derivative over the end region.
acc_logE	mean of peak-normalized logarithmic energy accelerative over the end region.
norm_pause	total pause durations normalized by the utterance duration.
syllarate	syllabic rate (number of syllables per second) normalized by the speaker's normal speaking tempo.
mean_norm_word_dur	mean of word duration which is normalized by the number of syllables the word has.
max_norm_word_dur	maximum of word duration which is normalized by the number of syllables the word has.
norm_utt_dur	utterance duration normalized by number of syllables in the utterance.

The features listed above are computed on approximately 300 utterances (181 confidence samples, 90 confusion samples, and 28 frustration samples) manually collected from the preliminary experiments on six subjects. The average statistics of the feature values are presented in Table 2 and figure 1.

Table 2. List of prosodic features (aside from the duration-related features) with respect to the three cognitive states

Feature	*Confidence*	*Confusion*	*Frustration*
F0_ratio	1.01	1.00	1.01
F0_reg_pen	0.32	-0.10	0.02
F0_reg_end	-0.29	0.48	-0.04
F0_norm	0.54	0.62	0.50
logE_ratio	0.9511	0.9029	1.0073
derive_logE	-0.0005	-0.0060	0.0013
acc_logE	0.0004	-0.0002	-0.0005
norm_pause	0.08	0.07	0.21
syllarate	0.94	0.95	0.81

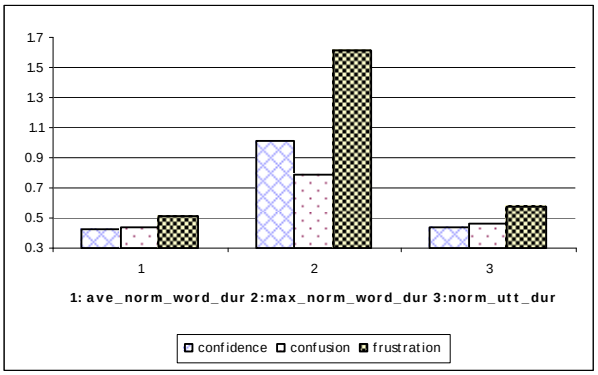

Figure 1. Duration-related feature plot with respect to the three cognitive states

In comparison of frustration state and the other two states, the duration-related features (see Figure 1) of frustration state are obviously larger, especially *max_norm_word_dur*. Meanwhile, the frustration state also has longer intra-sentence pauses, and lower syllable rate than the other two states.

For the energy-related features, *logE_ratio* is slightly more than 1.0 for frustration, whereas less than 1.0 for confidence and confusion, denoting remaining energy at the end of frustration, but decreasing energy at the end of confidence and confusion. The negative values of confidence and confusion and positive value of frustration in *derive_logE* illustrate the same argument. What *acc_logE* reveals is not clear.

For the pitch-related features, *F0_ratio* is almost the same for the three states. *F0_reg_pen* and *F0_reg_end* show the trend of tone variations in the near-end and end regions: rising and then falling for confidence; falling and then rising for confusion; slightly rising and then slightly falling for frustration. *F0_norm* indicates that the voiced region is highest for confusion and lowest for frustration.

4. Lexical information

Grammar defines different rules for different types of expression. For example, grammar has different rules for statements and questions. This is a reason why word information, especially initial words in a sentence is a clue for detection of intra-speaker variations. Another reason is that some key words have significant meaning and are critical for cognitive state detection. For example, "yes" almost definitely decides an utterance is confident. "I'm not sure" decides the user is in frustration. In this domain-specific dialogue system, 26 key words/phrases were selected as being most relevant to the problem of mental state determination. The key words/phrases are listed in the following table.

Table 3. A list of key words/phrases used in the mental state classification

yeah/yes	No	because
I think	Uhm	so
Which	Where	when
what/what's	I don't know	do you/I
can you/I/we	Could you	this one
is it	Why	these
I'm not sure	should I	how

The detection of key words/phrases makes use of a LVCSR system. Sampled at 11 KHz, the speech input is pre-emphasized and grouped into frames of 330 samples with a window shift of 110 samples. The speech signal is characterized by 13 MFCC components normalized by cepstral mean subtraction, and log-scaled energy normalized by the peak. Moreover, their deltas and delta-deltas are also computed. Therefore, each speech frame is represented by a vector of 42 features.

The key word/phrase spotting comprises detection, and subsequent verification to reduce false alarms. The detection portion is based on LVCSR, in which each key word/phrase is represented by the concatenation of the models of its component phones. Each subword model is a left-to-right 3-state HMM with 16 Gaussian mixtures per state. The universal background subword models trained from TIMIT database are further adapted to children of various ages and sex. The speech adaptation is based on maximum likelihood linear regression (MLLR) followed by maximum a Posteriori (MAP) adaptation. Recognition is accomplished by a frame synchronous Viterbi search algorithm to determine the sequence of words that maximizes the likelihood of a given utterance.

The putative results are further validated by employing a filler model, based on the geometric log likelihood mean proposed by Sukkar and Lee (1996). The keyword acceptance/rejection is determined by comparing V(O; W) with a predefined threshold, where

$$V(O;W) = \sum_{j=1}^{N} \{\log[L(O_j \mid s_j)] - \log[\frac{1}{M_j}\sum_{m=1}^{M_j} \exp(\gamma \log[L(O_j \mid s_j(m))])]^{1/\gamma}\}, \quad (1)$$

and the key word W is the concatenation of N subwords;

$L(O_j|s_j)$ is HMM likelihood score for O_j, the corresponding observation sequence, given s_j, the j^{th} subword model;

M_j is the total number of subwords in the corresponding cohort set of s_j;

$s_j(m)$ is the m^{th} subword in the corresponding cohort set of s_j; and

γ is constant.

5. System application

5.1. Overall structure

The detection of mental state is based on the prosodic features and key words/phrases associated with semantic meaning. The pilot study reveals that children have somewhat different expressions than adults when interacting with computers. One difference is that children prefer short sentences to long sentences for expression, probably due to the small vocabulary size and shortage of linguistic knowledge. For example kids, especially boys, like to express their agreement and denial using a simple "hum" but with different tones. In this case, prosody provides the key cue to discriminate the user's intention.

Nevertheless, some utterances are inherently ambiguous for mental state detection by means of prosody alone. For example, wh-questions, how-questions, and no-opinion statements (e.g., "I don't know") represent confusion, but may be uttered with prosody indistinguishable from the prosody of a confident utterance. Spotting some words/phrases embedded in the fluent speech helps discriminate the disparity and make a correct decision. The integration scheme of prosodic and lexical information is depicted in Figure 2.

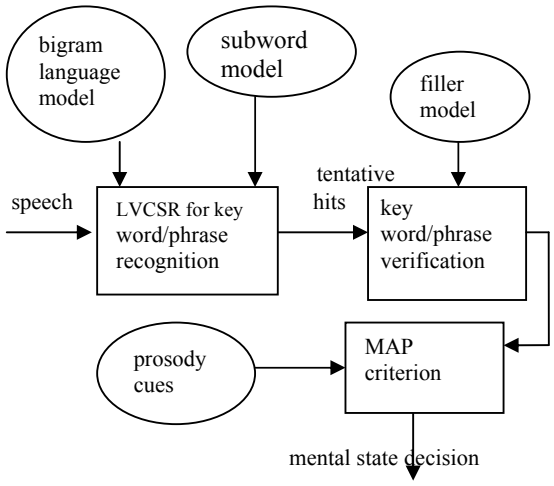

Figure 2. Overview of the mental state classifier

5.2. Uniqueness of children's speech

Children under 13 years old have very different acoustic characteristics depending on their age. The variability lies in two aspects: (1) age-dependent variability in terms of formants; (2) intra-speaker variability in terms of cepstral distance both within a token and across two repetitions (Narayanan and Potamianos, 2002). The vocal tracts of children are short and still growing. The shorter vocal tract length makes formant frequencies of children higher than those of adults. The variability results in degradation of ASR performance on children. It has been reported that the in-vocabulary word error rate for children is almost twice that of adult users. To compensate for these variabilities, we apply frequency warping to normalize vocal tract length (Zhan and Waibel, 1997). The scheme of frequency warping is shown in Figure 2. The frequency f is warped by means of a bilinear rule, which maps an input to an equal length of output in the frequency domain.

$$\varphi_{\beta_f}(f) = f + \frac{2f_N}{\pi}\tan^{-1}\left(\frac{(1-\beta_f)\sin(\frac{f}{f_N}\pi)}{1-(1-\beta_f)\cos(\frac{f}{f_N}\pi)}\right), \quad (2)$$

where f_N is the Nyquist frequency, and β_f is the frequency-dependent warping factor.

To compensate for the inter-speaker variability, different warping factors are used on groups of children

with the same age and sex. With reference to the data published in (Lee, et al., 1999), for each group, the warping factors at formants F1, F2 and F3 are computed as the ratio of average formant values of that group to those of adult males.

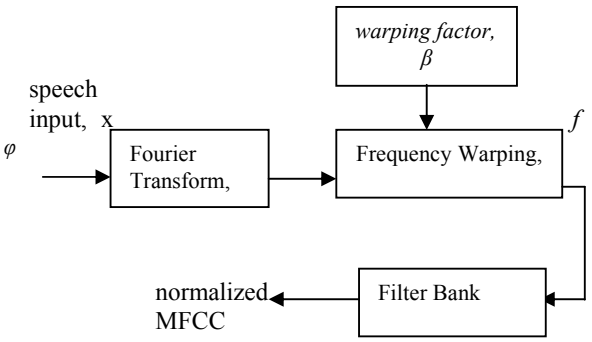

Figure 3. Vocal tract length normalization by frequency warping

To compensate for the intra-speaker variability, we put forward that the warping factors for frequencies other than the three formants are approximated by interpolation. The interpolation scheme is shown in Figure 3, where the three critical points are the group-dependent (F_1, β_1), (F_2, β_2) and (F_3, β_3), respectively.

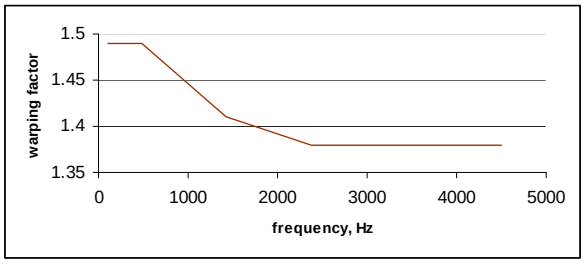

Figure 4. The interpolation pattern for deriving frequency-dependant warping factors to compensate intra-speaker variability

5.3. Classification strategy

Classification is divided into two steps. First is the classification of frustration and non-frustration. If the latter case is true, then further classify confidence and confusion.

The classification strategy is the integration of lexical and prosodic information by MAP. Denote the mental state as E, the recognized keywords or their combination set as W, and prosodic features as F. Then the a posteriori most probable hypothesized mental states, given the prosodic evidence and word identities, is derived by the MAP criterion

$$E^* = \arg\max_E P(E \mid W, F) = \arg\max_E P(W, F \mid E)P(E)$$

$$= \arg\max_E \{\log P(W \mid F, E) + \log P(F \mid E) + \log P(E)\}. \quad (3)$$

In the computation of P(W|F, E), insufficient training samples mean that some possible cases might not be observed during training. To compensate for the inherent sparseness of data, we used a standard unit-discount smoothing technique (Chen and Goodman, 1999). That is, each of the keywords is set to occur once more, in each context, than it really does:

$$P(W_i \mid F, E) = \frac{1 + T(W_i, F, E)}{\sum_{W_j} 1 + T(W_j, F, E)}, \quad (4)$$

where $T(W_i, F, E)$ is the number of times when W_i, *F and E* occur simultaneously.

6. Experiments

6.1. Experiment setup

Whereas our ultimate goal is oral communication between a computer agent and users, the preliminary experiments are carried out by a human tutor and the user subjects in the form of Wizard-of-Oz simulations. A user subject and the human tutor are sitting in two rooms. Wearing a head-set microphone, the user orally communicates with a computerized talking head, which is manipulated by the tutor and uses synthesized speech from typed text. The user's speech is transmitted through the microphone to a digital video-audio camera (SONY, model DCR-TRV240) placed at the right front of the user. The camera is connected with the earphone of the tutor sitting at the next room. Another function of the camera is to capture the facial expression of the user. The child user plays with the Legos in front of him/her during experiments. The activities of the user are captured by a digital camcorder (Panasonic), which is mounted on the top of the playing scene. Sometimes the tutor needs to demonstrate to the user how to play with the Lego set. For this purpose, a similar Panasonic camera is used on the tutor too. The demonstrating instruction by the tutor is transmitted through the digital camera to the user's computer and displayed on the screen.

In addition, an Infrared camera (CE, model EVI-D30) used for eye tracking is fixated opposite the user. A small transmitter is placed on the head of the user, and a receiver (Ascension Technology, model 6DF0B) is located at the back of the user to track the head motion of the user.

The duration of each experimental session varies from 20 to 35 minutes depending on the cooperative attitude of the user.

6.2. Human performance

Some children are talkative, and some are reticent. When the users are not sure how to respond to the

computer, some users will frankly express their frustration, some users speak in a hesitant tone, whereas some users just keep silent. We found the behaviour toward answering tough questions is different from child to child also. Some children are quick to respond, possibly in confusion or frustration mood; some children are thinking carefully, slowly responding, but with definite answers (confidence). Moreover, although confidence state more likely indicates correct answers, correct target responses are not always associated with confidence: confidence mood might have wrong answers, while frustration mood might have correct answers.

According to the utterances collected in the preliminary experiments so far, the ratio of confidence: confusion: frustration ≈ 0.6:0.3:0.1.

6.3. Experimental results and discussion

6.3.1. Experimental results

Twelve experiments were carried out so far. The test users are primary and early middle school students, aged from 7 to 13. Phoneme models were constructed using HTK, and trained using the TIMIT database. Models of children ages 7-9 were then adapted using the "CMU kids" database, a microphone speech data corpus distributed by the Linguistic Data Consortium that provides substantial dictation speech of children ages 7 to 9. Phoneme models for children of other ages had to be adapted using a few minutes of utterances by the users themselves. The models of prosody feature distribution were trained using approximately 300 sample utterances manually extracted from the data collected from six users. The tests were carried out on 80 utterances of another three users: two 12 year-old girls and one 11 year-old girl. Test results showed that ?% are correctly recognized. Test also showed that the non-stationary noise caused by Lego playing is a big error source.

6.3.2. Future work

Polzin and Waibel (2000) argue that the spectral information is also helpful for emotion recognition, because in the experiments of Lieberman and Michaels (1962), the emotion recognition performance degraded greatly if only pitch and energy are considered. We shall investigate the impact of spectral information on cognitive state detection when more data is available.

We notice that the dialogue context is potentially useful for cognitive state detection and needs to be investigated in the future. For example, when the user initiates a dialogue, then the user is almost always in a state of confidence or confusion. When the computer initiates a dialogue by a tough question, then confusion or frustration is more likely to occur than confidence.

7. Conclusion

Through the prosodic analysis and key word/phrase spotting of children's spoken language, the mental activities of children can be classified during their interaction with computers. The inter-speaker variability and intra-speaker variability of children's speech are compensated for by a vocal-tract-length-normalization (VTLN)-based technique. Tests on a set of 50 utterances collected from the project experiment showed the classification accuracy was ?%.

8. Acknowledgements

This work is supported by NSF grant number 0085980. Statements in this paper reflect the opinions and conclusions of the authors, and are not endorsed by the NSF.

9. References

Alter, K., Rank, E., Kotz, S. A., Toepel, U., Besson, M., Schirmer, A., and Friederici, A. D. 2003. Affective encoding in the speech signal and in event-related brain potentials. *Speech Communication, 40* (pp 61-70).

Batliner, A., Fischer, K., Huber, R., Spilker, J., and Nöth, E. 2003. How to find trouble in communication. *Speech Communication, 40* (pp. 117-143).

Batliner, A., Nöth, E., Buckow, J., Huber, R., Warnke, V., & Niemann, H. 2002. Duration features in prosodic classification: why normalization comes second, and what they really encode. *ISCA Workshop on Prosody in Speech Recognition and Understanding.*

Bell, L. and Gustafson, J. 1999. Repetition and its phonetic realizations: investigating a Swedish database of spontaneous computer directed speech. In *Proceedings of ICPHS99*. pp. 1221-1224.

Bosch, L. t. 2003. Emotions, speech and the ASR framework. *Speech Communication, 40* (pp. 213-215).

Chen, S. & Goodman, J. 1999. An empirical study of smoothing techniques for language modelling. In *Computer Speech and Language*, *13*, 359-394.

Edwards, J. 1991. The articulatory kinematics of final lengthening. In *the Journal of the Acoustical Society of America*. 89(1), 369-382.

Fernandez, R. and Picard, R. W. 2003. Modeling driver's speech under stress. *Speech Communication, 40* (pp. 145-159).

Galbraith, M. W. 1998. *Adult Learning Methods: a Guide for Effective Instruction.* Melbourne, FL: Krieger Publishing Company.

Graesser, A. C., Lehn, K. V., Rose, C. P., Jordan, P. W., & Harter, D. 2001. Intelligent tutoring system with conversational dialogue. In *AI Magazine*, *22*(4), 39-52.

Heeman, P. A., & Allen, J. F. 1999. Speech repairs, intonational phrases and discourse markers: modeling speakers' utterances in spoken dialogue. In *Association of Computational Linguistics*.

Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P., & Ess-Dykema, C. V. 1997. Switchboard discourse language modeling project final report. In *Johns Hopkins LVCSR Workshop*.

Kafai, Y., and Harel, I. 1991. Children learning through consulting: when mathematical ideas, knowledge of programming and design, and playful discourse are

intertwined. In *Constructionism*, Norwood, NJ: Ablex Publishing Corp.

Kafai, Y. & Resnick, M. 1996. *Constructionism in Practice: Designing, Thinking and Learning in a Digital World*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Kitazawa, S., Ichikawa, H., Kobayashi, S., & Nishinuma, Y. 1997. Extraction and representation rhythmic components of spontaneous speech. In *EUROSPEECH* (pp. 641-644).

Lee, S., Potamianos, A., & Narayanan, S. 1999. Acoustics of children's speech: developmental changes of temporal and spectral parameters. In t*he Journal of the Acoustical Society of America*, *105*(3), 1455-1468.

Levinson, S. E. 1985. Structural methods in automatic speech recognition. In *Proceedings of the IEEE*, *73*(11), 1625-1649.

Levinson, S. E. 1977. The effects of syntactic analysis on word recognition accuracy. In *Bell Syst. Tech. J.*, *57*, 1627-1644.

Levow, G. 1998. Characterizing and recognizing spoken corrections in human-computer dialogue. In Proceedings of COLING/ACL'98. pp. 736-742.

Li, Q., Zheng, J., Tsai, A., & Zhou, Q. 2002. Robust endpoint detection and energy normalization for real-time speech and speaker recognition. In *IEEE Trans. Speech Audio Processing*, *10*(3), 146-157.

Lieberman, P. & Michaels, S. B. 1962. Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. In *the Journal of the Acoustical Society of America*, *34*(7), 922-927.

Menezes, C., Pardo, B., Erickson, D., and Fujimura, O. 2003. Changes in syllable magnitude and timing due to repeated correction. *Speech Communication, 40* (pp. 71-85).

Miller, G. A., Heise, G. A., & Lichten, W. 1951. The intelligibility of speech as function of the context of the test materials. In *J. Exp. Psychol*., *41*, 329-335.

Morgan, N. & Fosler-Lussier, E. 1998. Combining multiple estimators of speaking rate. In *IEEE ICASSP*.

Narayanan, S. & Potamianos, A. 2002. Creating conversational interfaces for children. In *IEEE Trans. Speech and Audio Proc*., *10*(2), 65-78.

Polzin, T. S., & Waibel, 2000. A. Emotion-sensitive human-computer interfaces. In *ICSA Workshop on Speech and Emotion: a Conceptual Framework for Research*.

Schröder, M. 2003. Experimental study of affect bursts. *Speech Communication, 40* (pp. 99-116).

Shriberg, E., Bates, R. & Stolcke, A. A prosody-only decision-tree model for disfluency detection.

Shriberg, E. Stolcke, A. Hakkani-Tür, D. & Tür, G. 2000. Prosody-based automatic segmentation of speech into sentences and topics. In *Speech Communication 32(1-2): Special Issue on Accessing Information in Spoken Audio*.

Scherer, K. R. 2003. Vocal communication of emotion: a review of research paradigms. *Speech Communication, 40* (227-256).

Stevens, K. N. 1991. Sources of inter- and intra-speaker variability in the acoustic properties of speech sound. In *Proc. of the 12th International Congress of Phonetic Sciences*, 206-232.

Sukkar, R. A., & Lee, C.-H. 1996. Vocabulary independent discriminative utterance verification for non-keyword rejection in subword based speech recognition. In *IEEE Trans. Speech and Audio Proc.*, *4*, 420-429.

Tsukahara, W. & Ward, N. 2001. Responding to subtle, fleeting changes in the user's internal state. In *Proc. of the SIGCHI Conf. On Human Factors and Computing Systems*.

Wilensky, U. 1991. Abstract meditations on the concrete. In *Constructionism*, Norwood, NJ: Ablex Publishing Corp.

Zhan, P. & Waibel, A. 1997. Vocal tract length normalization for large vocabulary continuous speech recognition. In *CMU Computer Science Technical Reports*.