Children's Emotion Recognition in an Intelligent Tutoring Scenario

Tong Zhang, Mark Hasegawa-Johnson, and Stephen E. Levinson

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
{tzhang1, hasegawa, sel}@ifp.uiuc.edu

Abstract

This paper presents an approach to automatically recognize emotion which children exhibit in an intelligent tutoring system. Emotion recognition can assist the computer agent to adapt its tutorial strategies to improve the efficiency of knowledge transmission. In this study, we detect three emotional classes: c*onfidence*, *puzzle*, and *hesitation*. Emotion is detected by means of lexical, prosodic, spectral, and syntactic analyses of users' speech. An automatic speech recognition system serves as the fundamental constituent of the system. A robust classification and regression tree (CART) integrates the various information sources together for final decision. The effectiveness of the proposed approach has been tested on data collected by Wizard-of-Oz (WoZ) experiments. Our emotion recognition was speaker-independent, and yielded 91.3% accuracy. The test results showed that the spectral and duration-related prosodic features played very important roles in emotion recognition.

1. Introduction

We are developing an intelligent tutorial system (ITS) to provide a computer-based environment for education in science and technology, using a Lego construction set, with children of primary and early middle school ages. One important component of the system is emotion detection, i.e., detect emotion which children exhibit during their interaction with the computer. Through our WoZ experiments which will be described later, we find that children rarely exhibit strong emotions such as happiness and anger in a tutorial scenario. Rather, the frequently displayed emotion consists of confidence, puzzle, and hesitation. Recognition of these kinds of emotion is pragmatically useful, since it can not only help to develop a user-friendly interface but also help the computer agent to adapt its tutorial strategies to best meet the needs of students. When the student expresses his/her idea in a confident manner, e.g., "I saw the small one moving faster," the computer needs to evaluate the student's idea; and if the idea is correct, the computer will provide suggestions for the student's next action. When the student has questions, e.g., "what do you mean by spinning what?" the computer agent needs to answer his/her questions. When the student exhibits hesitation, 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, and get a better comprehension.

Emotion recognition in ITSs is being investigated in various ways. Emotion recognition can be achieved by sensing emotional-related physiological changes such as heart rate, breathing, blood pressure, and even muscle tightness [1]. Speech has also been used to automatically classify positive, negative, and neutral emotions in tutorial dialogues, and yielded 80.53% accuracy [2].

Speech is a rich information source for emotion recognition, and the speech-based emotion recognition is being widely investigated, although relatively little work has been done to apply this cutting-edge technology to ITS applications. Emotion recognition is based on various aspects of speech and language: (1) long-term prosody revealing phonetic variations is most commonly used for emotion recognition [3][4]; (2) short-term spectrum capturing vocal-tract movements [5]; (3) lexicon and syntax respectively associated with semantic meaning and linguistic structure [6][7]; and (4) dialogue context providing a knowledge source to help reveal emotion [8][9].

2. WoZ data corpus

2.1. Data collection

WoZ experiments allow speech data to be collected in a way much similar to a real computer-tutoring environment. In this study, the objective of the experiments is to help children learn some basic concepts of Mathematics and Physics through manipulating concrete objects (Legos) rather than solely handling abstract symbols [10]. The children are given gears of different sizes. The teeth on each gear are painted with different color pairs: red and blue, red and green, or blue and green. The tutor helps children by asking them questions, guiding them to use Legos to find solutions of the questions, and answering questions which they propose. Meanwhile, the tutor provides emotional support and consolation, and carefully adjusts his tutorial strategy according to emotion and learning progress of the children. For example, one question is about the ratio of teeth number and spinning cycles:

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? Why?

Children are expected to line up a 24-tooth gear and a 40-tooth gear along a beam and right next to each other, and then rotate the gears, counting and comparing the spinning cycles. Children should observe that gears with more teeth spin more slowly. Some children further discover that the product of teeth number and spinning cycles is the same for the two gears.

To date 29 experiments with 17 subjects have been carried out and transcribed, and we have collected 11.7hrs of audio-visual data. Many student subjects are silently (no talking) exploring Legos most of the time in the experiments. Some students even kept silent when the tutor tried conversation

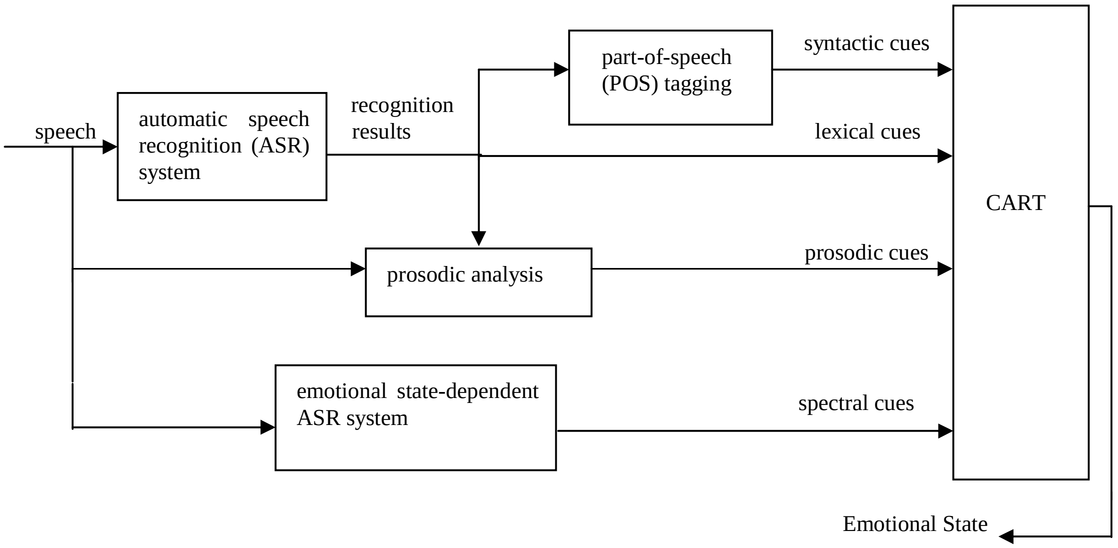

Figure 1: Overview of the emotional state classifier.

with them. However, some students were relatively easy to communicate and provided us relatively rich speech information for our study. In addition, some speech data were very noisy caused by Lego playing and heavy breath, etc., and had to be discarded. To date 714 students' utterances (students' turns) were used for this study, containing approximately 50mins of relatively clean speech. On average each utterance has 4.2s speech and 8.1 words.

2.2. Data annotation

The annotation criteria are listed in Table 1. Three annotators worked on the WoZ audio-video data independently of each other. The annotation was performed based on vision, speech content, and dialogue contexts. We used the Kappa statistics to evaluate consistency among the annotators, and yielded a score of 0.93, indicating a very good agreement.

Table 1: Emotional categories and their annotation criteria

Emotion	*Annotation Criterion*
confidence	Answer questions or explain his/her actions in fluent or slightly disfluent (such as occasional repair) speech; or send a command; or after finishing a task.
Puzzle	Ask questions such as yes/no questions and wh-questions; or key phrases such as “I don’t understand” and “I don’t know”. Question looks.
Hesitation	Answer questions or explain his/her actions in heavily disfluent and relatively slow speech; or key phrases such as “I’m not sure”. Grimace facial expression; slamming hands on table.

3. System description

Emotion recognition is an integration of the prosodic, lexical, spectral and syntactic analyses. The overall structure of the emotion classifier is depicted in Figure 1. A spontaneous continuous ASR system serves as the fundamental part of the classifier, because it provides information for the lexical and syntactic analyses, which are described in Sections 3.3 and 3.5, respectively. In addition, part of prosodic features used the recognition results as described in Section 3.2. An emotional state-dependent ASR system provides information for spectral analysis as described in Section 3.4. Finally, all information cues are integrated by a binary CART decision tree, which is described in Section 3.6, for emotional state classification.

3.1. Speech recognition

In the ASR, the speech signal is characterized by (MFCC + energy), Δ (MFCC + energy), and ΔΔ (MFCC + energy). The dimension of MFCC is 13. We use tied-state triphones, each of which is a left-to-right 3-state HMM with 32 Gaussian mixtures per state. The HMM is trained on the TIMIT data corpus. Children under 13 years old have very different acoustic characteristics from adults. We perform vocal-tract-length normalization with reference to adult male by frequency warping to compensate for children's short vocal tract. The detailed description on children's speech recognition is in [11]. The facility to reduce the first-pass search space and improve the recognition performance is realized by a bigram language model (BLM). Usually the construction of a BLM requires millions of words, but our WoZ data corpus is far less than the requirement. To make up for the data insufficiency, we propose an approach which derives from Switchboard transcriptions the POS-level linguistic information whereby to establish a word-level language model for our system. Specifically, the word-level bigram model is given by:

$$p(w_2 \mid w_1) = \sum_{POS_2, POS_1} \{p(POS_2, POS_1 \mid w_2, w_1)\, p(POS_2 \mid POS_1)\, p(w_2 \mid POS_2)\}$$

$$= \sum_{POS_1, POS_2} \{p(POS_2 \mid w_2)\, p(POS_1 \mid w_1)\, p(POS_2 \mid POS_1)\, p(w_2 \mid POS_2)\} \qquad (1)$$

where w_1 and w_2 are a pair of words in the vocabulary of the WoZ transcriptions, POS_1 and POS_2 are all possible part-of-speech tags of w_1 and w_2, respectively, $p(\mathrm{POS} \mid w)$ is the empirical frequency that a POS is attached to a given word w, and is computed from switchboard transcriptions, $p(\mathrm{POS}_2 \mid \mathrm{POS}_1)$ is the POS-level bigram probability, $p(w \mid \mathrm{POS})$ is the empirical frequency that a word occurs given a POS and is computed from the WoZ transcriptions. The POS-level BLM is generated from the POS-tagged switchboard transcriptions using the backoff smoothing technique.

3.2. Prosody analysis

The prosodic features are based on pitch, energy, pause, syllabic rate and duration. Table 2 lists the prosodic features used for emotion recognition. The derivation of pitch, energy, pause, syllabic rate, and duration is described in [11].

Table 2: List of the prosodic features

Feature	*Description*
f0_ratio	Ratio of mean *f*0 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_cubic_ratio	Ratio of mean *f*0 over the end region and the penultimate region after cubic interpolation.
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.
acce_logE	Mean of peak-normalized logarithmic energy acceleration over the end region.
norm_pause	Total pause durations normalized by the utterance duration.
syllarate	Syllabic rate 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 in that word.
utt_dur_by_syllable	Utterance duration normalized by the number of syllables in that utterance.
utt_dur_by_word	Utterance duration normalized by the number of words in that utterance.
max_abs_word_dur	Maximum of absolute word duration.

3.3. Lexical analysis

Psychology researchers found that certain words were associated with emotion expression [7]. We manually generate a list of approximately 50 key words/phrases which are emotionally salient by analyzing the data corpus. Furthermore, to handle the problem of data sparcity, we define a few key lexicon classes to cluster the key words/phrases which exhibit similar lexical behavior. The key lexicon classes consist of (1) affirm (yes, yeah, yep, no); (2) digit (one, two, three, etc.); (3) known (I think, I know); (4) uhm (ahh, um, ahmm, etc); (5) reason (because, so); (6) unknown (don't know, don't understand, etc.); (7) auxiliary (can you, would you, should we, can I, etc.); (8) unsure (not sure, not exactly sure); and (9) wh-word (which, when, how, etc.). The spotted key word(s)/phrase(s) from a recognized word string is (are) assigned to the corresponding lexicon class(es), which is (are) considered as the lexical cue.

3.4. Spectral analysis

The only difference between the emotional state-dependent ASR system and the routine ASR system is that the phoneme models of the former system are adapted to data which are categorized into emotional classes. For example, the confidence-dependent phoneme model is adapted using confidence utterances in the data corpus. For a given utterance, ASR based on each emotional state-dependent model generates a word string associated with an acoustic likelihood score. We compare the corresponding acoustic likelihood scores of the three different phoneme models, and use the differences as the spectral cues. The spectral features are listed in Table 3.

Table 3: List of the spectral features

Feature	*Description*
spect_confid-puzzle	acoustic likelihood score using the confidence-dependent model minus acoustic likelihood score using the puzzle-dependent model
spect_confid-hesit	acoustic likelihood score using the confidence-dependent model minus acoustic likelihood score using the hesitation-dependent model
spect_puzzle-hesit	acoustic likelihood score using the puzzle-dependent model minus acoustic likelihood score using the hesitation-dependent model

3.5. POS tagging

In this study, the syntactic composition of sentences is useful to reveal the three emotional states. The syntactic composition pattern of the beginning region in statements (often denotes confidence) and questions (usually denotes puzzle) are quite different. Sentences elicited by a hesitated speaker are usually out-of-grammar and incomplete. The final word, or the exit of a sentence, is especially useful in detecting the completeness/incompleteness of a sentence. For example, the exit word with a conjunction POS usually indicates that the sentence is incomplete. The context of the exit word usage is also important. For example,

This gear is a much bigger
This gear is much bigger

The first sentence is incomplete while the second sentence can be considered complete, although they have the same exit word.

Spoken language usually has loose-grammar structure and is filled with dysfluencies such as repair and repetition. Therefore, the POS string of an entire spoken sentence is not an efficient feature. In stead, we use the POS of the first 3 words and the last 3 words of a sentence as the syntactic cues for emotion recognition. By limiting the analysis object to local regions, the opportunity of getting loose-grammar and dysfluency is greatly reduced. The POS tagging is automatically performed by Roth's tagger [12].

3.6. Information fusion

The various information sources are integrated probabilistically via a binary CART, which automatically uncovers hidden structure of training data to generate predictive models for classifications [13]. Each variable in the data vector $\mathbf{x}=(x_1, x_2 \ldots x_n)$ is in discrete or character type. The CART introduces the idea that over-grows trees and then pruns back to ensure important structure not to be overlooked by stopping too soon. Gini, a criterion for splitting a single variable, is used for growing the binary tree.

4. Results and Discussion

The speech data contained 714 user's turns, of which 441 were *confidence*, 216 were *puzzle*, and 57 were *hesitation*. In CART learning and classification, prior probabilities matched total sample frequencies, i.e., *confidence* : *puzzle* : *hesitation* = 0.62 : 0.30 : 0.08. We used the 10-fold cross-validation method for system evaluation, and recorded the most accurate result regardless of the tree size. Our study showed that for syntactic analysis, processing only the last word (exit word) yielded slightly better results than processing the last three words in an utterance. We recorded the result in the former case, which was 91.3% accuracy. A confusion matrix summarizing the classification performance is shown in Table 4. We compute precision, recall, and *F*-score for each of the three emotional states, and present the result in Table 5.

Our study showed the spectrum and duration-related features are the most important variables for emotion recognition. Table 6 lists the seven variables which are most important in emotion recognition. The score reflects the contribution each variable makes to the classification performance [13].

Table 4: Confusion matrix of emotion classification

	confidence	*puzzle*	*hesitation*
confidence	421	15	5
puzzle	23	189	4
hesitation	9	6	42

Table 5: Precision, recall and F-score

Emotion	*Precision*	*Recall*	*F-score*
confidence	0.929	0.955	0.942
puzzle	0.900	0.875	0.887
hesitation	0.824	0.737	0.778

Table 6: Variable importance ranking

	Feature	*Score*
1	Spect_confid-puzzle	100%
2	max_norm_word_dur	88.46%
3	max_abs_word_dur	85.52%
4	Spect_confid-hesit	36.96%
5	utt_dur_by_syllable	30.90%
6	Spect_puzzle-hesit	29.94%
7	mean_norm_word_dur	28.78%

5. Conclusion

We have addressed an approach to automatically recognize children's emotion in an intelligent tutoring scenario. Emotion recognition using speech was accomplished by various types of speech analyses: lexicon, prosody, spectrum, and syntax. All of the analytical results were integrated by a robust binary CART. In addition, an automatic speech recognition system served as the fundamental constituent of the emotion recognition system. We proposed an approach which derived from the Switchboard transcriptions the POS-level linguistic information whereby to establish a word-level language model for our system. Our speaker-independent emotion recognition yielded an average of 91.3% accuracy. Our test results also showed that the spectral and duration-related prosodic features were very important for emotion recognition.

6. References

[1] Reynolds, C. and Picard, R. W. "Designing for affective interaction," *The 9th Intl. Conf. on Human-Computer Interaction*, New Orleans, Louisiana, 2001.

[2] Literman, D. and Forbes, K. "Recognizing emotions from student speech in tutoring dialogues," *The IEEE Automatic Speech Recognition and Understanding Workshop*, St. Thomas, Virgin Islands, 2003.

[3] Cowie, R. and Douglas-Cowie, E. "Automatic statistical analysis of the signal and prosodic signs of emotion in speech," *ICSLP*, 1996.

[4] Fernandez, R. and Picard, R. W. "Modeling drivers' speech under stress," *Speech Communciation*, 40: 145-159, 2003.

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

[6] Batliner, A., Huber, R., Niemann, H., Nöth, E., Spilker, J., and Fischer, K. "The recognition of emotion," In W. Wahlster (ed.) *Verbmobil: Foundations of Speech-to-Speech Translations*, pp122-130, Springer, New York, Berlin, 2000.

[7] Plutchik, R. *The Psychology and Biology of Emotion*, Harper-Collins College, New York, NY, 1994.

[8] Batliner, A., Fischer, K., Huber, R., Spilker, J., and Nöth, E. "How to find trouble in communication," *Speech Communication*, 40: 117-143, 2003.

[9] Ang, J., Dhillon, R., Krupski, A., Shriberg, E., and Stolcke, A. "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," *ICSLP*, Denver, Co, 2002.

[10] Wilensky, U. "Abstract meditations on the concrete," (I. Harel & S. Papert, ed.) *Constructionism*, Ablex, Norwood, NJ, 1991.

[11] Zhang, T., Hasegawa-Johnson, M., and Levinson, S. E. "Mental state detection of dialogue system users via spoken language," *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, Japan, 2003.

[12] M. Munoz, V. Punyakanok, D. Roth and D. Zimak. 1999. A learning approach to shallow parsing. *EMNLP-WVLC'99*.

[13] Salford Systems. 2002. *CART for Windows User's Guide*. www.salford-systems.com