Combining Language and Speech Features to Predict Students' Emotions in E-Learning Environments

Liang-Chih YU*, Shou-Fang LIANG, Wei-Hua LIN
Department of Information Management, Yuan Ze University, Taiwan, R.O.C.
*lcyu@saturn.yzu.edu.tw

Abstract: Emotions play an important role in e-learning environments. Text and speech have been recognized as convenient and natural means for expressing emotions, and are increasingly used in human-computer interaction interfaces for e-learning applications, indicating that language and speech could potentially be used to predict learner emotions. In this study, we investigate the use of speech and language features for automatic emotion recognition. A corpus of emotion-laden sentences was collected from student-teacher dialogs in the context of mathematics instruction. The corpus was then annotated to analyze emotion types as they occurred in e-learning applications. The speech and language features were then used to build several classifiers for emotion recognition. Experiments show that the two features combined yielded better results than either feature alone. In addition, among speech features, energy and formant are found to best contribute to successful classification.

Keywords: Emotion recognition, natural language processing, speech processing

Introduction

Students frequently react to satisfactory or dissatisfactory learning performance by expressing positive or negative emotions, which, in turn, may have an impact on subsequent learning outcomes [1][2]. For instance, Rodrigo et al suggested that boredom may have a negative impact on student achievement, while confusion may have both positive and negative effects [2]. This has raised interest in technological solutions for automatic emotion recognition because accurately assessing changes in learner emotional states can allow e-learning systems to provide appropriate suggestions, thus improving learning outcomes.

Text and speech have been recognized as convenient and natural means for expressing emotions, and are increasingly used in human-computer interaction interfaces for e-learning applications such as computer supported collaborative learning (CSCL) [3][4] and intelligent tutoring systems [5][6]. For example, text-based synchronous online chat can be used for group discussion to support collaborative learning [3]. Asynchronous online discussion forums also facilitate knowledge sharing through posting and reading forum articles [4]. Speech has been integrated to help students interact with intelligent tutoring systems [5][6]. This increasing use of text- and speech-based interfaces positions both language and speech as potential features for identifying learner emotions in e-learning applications. Previous research has also demonstrated the effectiveness of using language and speech features for emotion recognition, but mainly in non-e-learning domains such as identifying positive and negative emotions (binary) [7], six basic human emotions [8], and specific emotion types in business [9][10] and medical domains [11][12]. Very little research has investigated the use of language or speech features for emotion recognition in e-learning applications [7]. In addition to language and speech features, mouse movements, facial features and body posture have also been investigated for identifying learner emotions [13][14].

Table 1. Corpus annotation results.

Emotion	Num. of sentence	Proportion	A1-A2 Agreement	A1 Accuracy	A2 Accuracy
Delight	194	26%	93.30%	94.85%	98.45%
Contempt	53	7%	62.26%	77.36%	84.91%
Boredom	81	11%	69.14%	88.89%	80.25%
Frustration	99	13%	59.60%	79.80%	78.79%
Confusion	134	18%	84.33%	92.54%	91.79%
Others	198	26%	87.88%	89.90%	97.98%
Sum/Avg.	759	100%	81.16%	89.33%	91.70%

In this paper, we investigate the use of both speech and language features to identify student emotions. To this end, we first collected a corpus of emotion-laden sentences from student-teacher dialogs in the context of mathematics instruction. The corpus was then annotated to analyze various emotion types as they occurred during use of e-learning applications. Finally, the speech and language features were combined to build several classifiers for emotion classification.

1. Corpus Annotation and Analysis

1.1 Corpus Annotation

The corpus collection process involved communication among three mathematics teachers and 149 students in discussing mathematical problems in the classroom. A total of 759 sentences were collected from student-teacher dialogs to form an emotion text corpus, with emotion types classified as Delight, Contempt, Boredom, Frustration, and Confusion. Sentences in the corpus not explicitly characterized by a specific emotion type were categorized as Other.

To analyze student emotions, the three mathematics teachers annotated the corpus to create a standard of the various emotion types. Each sentence in the corpus was first annotated with one of the six emotion types (including Other) by two teachers (annotators). In case of disagreement between the two annotators, the disputed sentence was judged by the third teacher (adjudicator) for a final decision. Post-adjudication proportions of the various emotion types and the accuracy of the two annotators could then be calculated from the corpus. The annotation results presented in Table 1 show that around 74% of the sentences in the corpus contained an emotion type, while the remaining 26% were out-of-domain sentences (i.e., “Other”). Among the five emotion types, Delight and Confusion were found to predominate.

Table 1 shows that the annotators A1 and A2 agreed on 81.16% of the sentences reviewed. Agreement regarding Contempt and Frustration was relative low, indicating that these two emotion types were more ambiguous. For example, Contempt may be misclassified as Delight, while Frustration may be misclassified as Boredom or Confusion. The accuracy of A1 and A2 (as calculated by their consistency with the adjudicator for samples for which there was disagreement) was 89.33% and 91.70%, respectively. Such human (expert) results can be viewed as the upper bound for automatic emotion classification using machine learning algorithms. The accuracy for Frustration for both annotators was relatively low, again indicating that this emotion type was more ambiguous.

Table 2. Linguistic features and sample sentences for the emotion types.

Emotion	Example sentence	Linguistic feature
Delight	I made a big progress this time. Oh! Great! This question is so easy.	progress, great, easy, simple
Contempt	This question is so stupid and deserves no response. This question is too elementary. Even a kid can do it.	stupid, basic elementary, kid
Boredom	That's so bored. I have addressed such kind of questions many times before. I don't want to waste my time on such a tedious question.	bored, boring, tedious
Frustration	That's too bad. I will be failed. Forget it. That's too hard.	bad, fail, hard, difficult
Confusion	This question is ambiguous. I do not understand the meaning. Why the question can be solved in this way?	ambiguous, why, weird, confuse,

Table 3. Prosodic features for each emotion types.

	Delight	Contempt	Boredom	Frustration	Confusion
Pitch Mean	increased	normal or increased	decreased	decreased	increased
Pitch Max	increased	increased	increased	decreased	increased
Pitch Min	increased	decreased	decreased	decreased	decreased
Energy Mean	increased	normal	increased	decreased	decreased
Energy Max	increased	increased	increased	decreased	decreased
Energy Min	increased	decreased	increased	decreased	decreased
Formant Mean	f1,f5 increased; f2-f4 decreased	f1,f3-f5 increased; f2 decreased	f1,f5 increased; f2-f4 decreased	f1,f3,f5 increased; f2,f4 decreased	f1,f2,f4 decreased; f3,f5 increased
Formant Max	f1-f3 increased; f4,f5 decreased	f1-f5 increased	f1-f4 decreased; f5 increased	f1,f3-f5 increased; f2 decreased	f1-f5 increased
Formant Min	f1,f4-f5 increased; f2,f3 decreased	f1,f3 decreased; f2,f4,f5 increased	f1-f5 increased	f1 decreased; f2-f5 increased	f1,f2 increased; f3-f5 decreased

1.2 Linguistic Features

Table 2 presents several sample sentences for each of the five emotion types. Students may express Delight when they are satisfied with their learning performance or when facing easy questions, but may express Contempt if the questions are too simple. Students may also express Boredom if they feel the questions are pointless or senseless. Conversely, students may express Frustration when they are worried about their performance or when facing difficult, and Confusion when facing ambiguous or incomplete questions. The last column in Table 2 summarizes a number of linguistic features for the various emotion types.

1.3 Speech Features

A total of 379 sentences were randomly selected for recording. The input waveforms were captured at 16kHz, a frame length of 33ms and an average length of utterance 3 seconds.

Table 4. Classification accuracy of different methods with different features (% accuracy).

	Two-class			Five-class		
	NB	C4.5	SVM	NB	C4.5	SVM
Language	83.64	74.41	89.45	64.38	57.26	70.45
Speech	70.45	77.84	76.52	38.26	51.98	55.67
Language + Speech (All)	85.22	79.16	91.29	67.02	59.63	72.03
Language + Pitch	81.79	73.35	88.92	62.80	56.73	69.92
Language + Energy	85.49	**84.17**	89.45	63.59	59.63	69.66
Language + Formant	81.05	74.14	90.50	65.17	55.41	72.30
Language + Pitch + Energy	**86.28**	83.11	89.71	65.17	58.84	70.18
Language + Pitch + Formant	81.27	75.46	90.50	64.12	55.94	72.03
Language+Energy+Formant	86.02	79.68	**91.56**	**67.28**	**60.69**	**72.56**

Recording was conducted in an office environment without obtrusive background noise. To ensure the quality of the recorded corpus, objective tests were performed to validate the correctness of the recorded data which was evaluated by averaging responses from all test subjects. The ground truth of most utterances was decided by a unanimous vote, thus giving the selected utterances significance.

Table 3 summarizes the analysis for each prosodic feature with respect to the various emotion types. According to our observations, the energy related features (i.e., mean, max and min) are useful for differentiating between high and low active states such as Delight and Frustration. The pitch related features are useful for discriminating between both Frustration and Confusion, and Delight and Boredom. In addition, the formant is also an important feature for discriminating among the various emotion types.

2. Experimental Results

The classifiers used in this study include the Support Vector Machine (SVM), C4.5, and the Naïve Bayes (NB) classifier from the Weka Package [15][16]. Each classifier was trained using language features (i.e., individual words), speech features (i.e., pitch, energy and formant as in Table 3), and both. A total of 379 recorded utterances were analyzed with 10-fold cross-validation. Each test utterance was classified as belonging one of the five emotion types from Table 2. A two-class classification was also performed by dividing the five emotion types into positive (Delight and Contempt) and negative emotions (Boredom, Frustration, and Confusion). Performance is measured as a function of *accuracy*, i.e., the number of correctly classified utterances divided by the total number of test utterances.

Table 4 shows the results of different classifiers with different features. For all classifier in both two-class and five-class classification, combining the speech and language features is found to yield higher performance than either individual feature. In addition, different features made different contributions to different classifiers. For NB and C4.5, Energy was the most promising feature because Energy-related feature combinations (i.e., Language+Energy, Language+Pitch+Energy, and Language+Energy+Formant) were more accurate than the other combinations. Conversely, for SVM Formant was found to be the most promising feature. The highest accuracies for two-class and five-class classification were 91.56% and 72.56%, respectively, indicating that there is still much room for improvement in five-class emotion classification.

3. Conclusion and Future Work

Speech and language features are used to identify emotions from a corpus of learner utterances collected within the context of mathematics instruction. The corpus is analyzed to determine emotion types, along with their associated speech and language features. Experimental results show that combining the two features yielded higher performance than using either feature alone and, among the speech features, energy and formant were found to make the greatest contribution to accurate identification. Future work will investigate other significant features to further improve classification performance. An additional possible direction is to realize emotion recognition in text and speech based e-learning applications.

Acknowledgements

This work was supported by the National Science Council, Taiwan, ROC, under Grant No. NSC100-2632-S-155-001 and NSC99-2221-E-155-036-MY3. The authors would like to thank the anonymous reviewers for their constructive comments.

References

[1] D'Mello, S., Graesser, A. & Picard, R. W. (2007). Toward an Affect-Sensitive AutoTutor. *IEEE Intelligent Systems*, 22(4), 53-61.

[2] Rodrigo, M. M. T., Baker, R. S. J. D. & Nabos, J. Q. (2010). The Relationships between Sequences of Affective States and Learner Achievement, *Proceedings of ICCE'10* (pp. 56-60), Putrajaya, Malaysia.

[3] Chiu, C. H. & Hsiao, H. F. (2010). Group Differences in Computer Supported Collaborative Learning: Evidence from Patterns of Taiwanese Students' Online Communication. *Computers & Education*, 54(2), 427-435.

[4] Yeh, Y. C. (2010). Analyzing Online Behaviors, Roles, and Learning Communities via Online Discussions. *Educational Technology & Society*, 13(1), 140-151.

[5] Litman, D. J. & Silliman, S. (2004). ITSPOKE: An Intelligent Tutoring Spoken Dialogue System, *Proceedings of HLT/NAACL'04*, Boston, MA.

[6] Ros´e, C. P., Litman, D., Bhembe, D., Forbes, K., Silliman, S., Srivastava, R. & K. VanLehn, (2003). A Comparison of Tutor and Student Behavior in Speech Versus Text Based Tutoring. *Proceedings of the Workshop on Building Educational Applications Using Natural Language Processing at HLT/NAACL-03*, Edmonton, Canada.

[7] Lee, C. M. & Narayanan, S. S. (2005). Toward Detecting Emotions in Spoken Dialogs. *IEEE Trans. Audio, Speech, and Language Processing*, 13(2), 293-303.

[8] Cook, N. D., Fujisawa, T. X. & Takami, K. (2006). Evaluation of the Affective Valence of Speech Using Pitch Substructure. *IEEE Trans. Audio, Speech, and Language Processing*, 14(1), 142-151.

[9] Bai, X. (2011). Predicting Consumer Sentiments from Online Text. *Decision Support Systems*, 50(4), 732-742.

[10] Xu, K., Liao, S. S., Li, J. & Song, Y. (2011). Mining Comparative Opinions from Customer Reviews for Competitive Intelligence. *Decision Support Systems*, 50(4), 743-754.

[11] Yu, L. C., Chan, C. L., Lin, C. C. & Lin, I. C. (2011). Mining Association Language Patterns Using a Distributional Semantic Model for Negative Life Event Classification. *Journal of Biomedical Informatics*, 44(4), 509-518.

[12] Wu, C. H., Yu, L. C. & Jang, F. L. (2005). Using Semantic Dependencies to Mine Depressive Symptoms from Consultation Records. *IEEE Intelligent Systems*, 20(6), 50-58.

[13] Horiguchi, Y., Kojima K. & Matsui, T. (2009). A Study for Exploration of Relationships between Behaviors and Mental States of Learners for an Automatic Estimation System, *Proceedings of ICCE'09* (pp. 173-175), Hong Kong, China.

[14] Graesser, A. C., Chipman, P., Haynes, B. C. & Olney, A. (2005). AutoTutor: An Intelligent Tutoring System with Mixed-Initiative Dialogue, *IEEE Trans. Education*, 48(4), 612-618.

[15] Witten, I. H. & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufmann, San Francisco.

[16] Zhang, P., Zhu, X., Shi, Y., Guo, Li. & Wu, X. (2011). Robust Ensemble Learning for Mining Noisy Data Streams. *Decision Support Systems*, 50(2), 469-479.