Modeling Affective States of Students to Aid Intelligent Tutoring Systems

Alekhya Gumidelli*, Divya Nalam, Divya Ramesh,
Louis-Phillipe Morency, and Stefan Scherer

University of Southern California
Los Angeles, California - 90007
{gumidell,dnalam,dramesh}@usc.edu
{morency,scherer}@ict.usc.edu

Abstract. Intelligent Tutoring Systems (ITS) could be a potential solution to many issues related to education. They can be used to reduce student-teacher ratio providing better individual attention to students, make education available in remote and underdeveloped locations, promote open-and free education for all. However, the biggest challenge for an ITS is to be able react to students in the same way as a human tutor does. This involves spontaneously adapting to the students' understanding levels and having to take decisions as to either elaborate on a concept, explain the same concept using different words, or to move on to the next topic. In our study, we have explored the possibility of using the audio-visual cues of students to detect their understanding with respect to the response of the human tutor. The analysis has been in a real-classroom scenario unlike acted datasets that are generally used in such studies. We have been able to achieve an accuracy of 62% on this dataset in classifying the students' levels of understanding which have been compared to ground truth labels of annotation derived by a majority vote from 3 annotators. These results are quite promising and indicate that using multimodal cues from a real-classroom scenario helps to better model the responses of an ITS in a student-ITS interaction.

Keywords: Multimodal, Intelligent Tutoring Systems, Affective Computing, audio-visual cues

1 Introduction

The aim of this project is to determine the understanding and the involvement of a student in an interactive student teacher setup in order to help improve Intelligent Tutoring Systems (ITS). Student engagement is vital to learning, and it helps if teachers are able to perceive the feedback from students that they often receive in terms of facial expressions, gaze and body postures. There is no point in the teacher going ahead if the student is not able to cope with the teacher. Traditionally, an ITS uses the answers to the questions posed to

* Corresponding Author

the students as a measure of understanding of students. It is not possible to keep the frequency of such questions too high. Thus, many a times the students can guess the right answers, or trick the system and move ahead. However, in an interactive student-tutor environment, the non-verbal cues often reflect on the difficulty, clarity and quality of the lecture content being presented. This information gives the ITS a better understanding of the students. Non-verbal cues give another dimension to the verbal answer given by the student. These gestures help us classify into distinct labels, thus giving a deeper understanding on the state of the student. By determining whether the student has understood or not the ITS can choose to either repeat a topic in a more elaborate way or go faster accordingly. Since the goal of an ITS is to replicate a human tutor as closely as possible in terms of receiving emotions, the expression of students to an ITS should be similar to that of an ITS. Due to non availability of a responsive ITS, and since the goal is to try and get as close to possible as a real teacher, a real student teacher interaction has to be considered. Thus in this study, we have considered a real classroom scenario where a professor interacts with a set of students. We describe the study in detail in the following order: Section 2 describes the previous work in this area, Section 3 details the dataset used in the study, Section 4 describes the qualitative analysis before carrying out the experiments. The experimenal setup and results are discussed in Sections 5 and 6 respectively. We conclude this study with Section 7.

2 Previous Work

Joseph et al[1] did a study where the human tutors interacted through a text based interface and the self-reports of students were used as ground truth labels. They compared unimodal, bimodal and multimodal features. For annotation, they followed the dialogue act annotation scheme as given in[2]. CERT[3] was used for facial expression recognition which detects faces, finds features for the nearest face that is detected and outputs weights for each facial action that is detected. In this paper, since it was a text based interaction, the audio modality was not taken into consideration. D'mello and Graesser[4] combined conversations, body gestures and facial expressions in their study. They used feature level fusion. The user study was based on 28 learners who did a 32 minute tutorial session with a virtual tutor. The data was annotated according to six affective states. However, practical applications require the detection of naturalistic emotional expressions that are grounded in the context of the interaction. Sarrafzadeh et al[5] developed an Affective Tutoring System (ATS) which changes its tutoring style based on the affective states of the student. The facial expressions along with gestures were used to detect the affective state of the student. The history is updated classifying the student's response to the question and using expression information. This generates a set of weighted recommendations for the ATS's next action. An in-house facial expression analysis system and gesture recognizer, which was implemented using a multilayer feed-forward Artificial Neural Network, were used. This again did not take into consideration

the audio cues of the student. The labels considered were boredom, confusion, inattention and anxiety which are what we are considering in this study as well.

3 Dataset/UserStudy

The dataset comprises of the video and audio recordings of a set of 7-8 students discussing with a professor seated around a round table, a set of senior/graduate lessons on Modern & Contemporary American Poetry. These video recordings were downloaded from the course page on Coursera[6]. The videos consist of the professor explaining a part of the poem for a brief time and then directing a question to each of the students. For example, the question could be "What could the poet mean when he says such there are more windows than doors?" These type of questions elicit responses where the student explains his views in two-three sentences, sometimes in agreement with the professor, and sometimes not. There are quite a few times when the professor points out that the views of the student are incorrect, and also appreciates the student if a new view is expressed. Thus, it is in an interactive environment that is quite different from the datasets used in the previous studies. Out of the set of video lectures, 5 weeks of lectures have been considered. We have based our study on the cues exhibited by the students while they are listening and answering questions posed to them. The setting includes a camera which is moved around and focused on the student at whom the question is directed. Snapshots from the dataset can be seen in Figs. 1a 1b, 1c and 1d.

(a) (b)

(c) (d)

Fig. 1: Snapshots of frames from our dataset

3.1 Preprocessing

The entire set of videos could not be used by itself as there were many parts in the videos where the automatic video annotation tools would fail. Since it was a single camera which was rotated to focus on the speaker, it led to swift camera movements, multiple people on the screen, especially when the professor is explaining a concept and the students are listening. The videos also consisted clips which were profile views of the speaker. Since none of these clips can be used, preprocessing was required. The steps involved in this are as follows:

1. The clippings which contained frontal face and slight profile views were retained. The rest were rejected.
2. In clips that contained multiple faces, only one face was considered based on the following conditions:
   - The person was the main speaker of the clipping.
   - The person was the intended listener i.e., The professor was explaining to that student, or asking the student a question.
   - The person was most prominent on the screen at that instant.

After removing unwanted frames, clips ranging from 4 minutes to 7 minutes were obtained from videos that were originally 15-20 minutes long. We collected 15 such clips having about 100 minutes of data in all.

3.2 Segmentation

The segmentation and annotations were carried out using a language annotation tool called ELAN[7]. The segmentation was at an utterance level. Every new sentence(meaningful phrase) uttered by a person(mostly students) was considered a new utterance. Turn taking was also taken into consideration. When a there was a change in speakers, it was considered a new utterance. This utterance level of segmentation resulted in about 50 - 60 segments from each clipping which were under 10 seconds each. There were a few exceptions in which the segments were longer, and ran up to 15 seconds.

3.3 Annotation

The authors each individually annotated each segments based on a prior agreement of four levels of affective states as given below:

1. Confused
2. Neutral
3. Well understood
4. Listening

The inter-rater agreeability calculated using Krippendorf's Alpha resulted in a high value of 0.766032 from the three annotators, thus indicating a high degree of agreement in the affective states defined above.

Due to limited availability of listening data, it was impossible to classify the listening states under the labels listed above. Hence, all of the listening segments were annotated with a label '4' so as to keep those clips from influencing the classification of the other three labels defined for different levels of understanding.

The majority vote derived from these annotations were then considered as the ground truth labels to the classifiers.

We obtained 592 segments in all with 9.97% of the data assigned the confused label, 38.68% under neutral, 20.27% under well-understood and the remaining 31.08% annotated as listening.

4 Qualitative Analysis

In the process of annotating our dataset, we observed that different combinations of audio-visual cues gave us different impressions about the students . Also, the presence of the same cue may be interpreted differently based on whether the student was listening or speaking while displaying it. Following are some observations specific to the labels we have chosen during annotation:

1. **Confused**: When a student is total confused during their turn to speak, we noticed that their pauses were accompanied by a wide smile (probably an effort to fill awkward silence). If they are confused and do speak, we noticed that the speech was accompanied by laughter. We also noticed that sometimes when the students are slightly confused, they are hesitant to talk. This is probably an indication that the student has not yet gained complete understanding of the subject and is thinking hard to choose words carefully. This is characterized by frequent-short pauses/stammering accompanied by gaze away from the intended listener (in this case, the professor). This can be seen in Fig. 3a.
2. **Neutral**: While giving straight-forward answers or expressing general opinions, students mostly adopted a passive expression and monotonous speech. We noticed wide variations in the baselines of the students (some of them were very animated while others showed subtle expressions). However, when the student does not have very strong opinions about a topic, the audio/visual dynamics displayed by the student do not change much within an utterance. This can be seen in Fig. 2b.
3. **Well understood**: While expressing strong opinions, students tend to stress on specific words. There were some instances where the students showed high confidence in their answers. This was characterized by high energy levels, high frequency and faster speech compared to their baseline. Sometimes, confidence was also expressed with eye-brow raise movement and head nods while speaking. This can be seen in Fig. 2c.
4. **Listening**: As the video has the professors voice embedded while the students are listening, the classifier takes into account this audio. Hence, the movement of the lips was taken into consideration in this case. A snapshot of a student that is listening is shown in Fig. 2d.

(a) Confused expression (b) Neutral expression

(c) Confident expression (d) Listening state

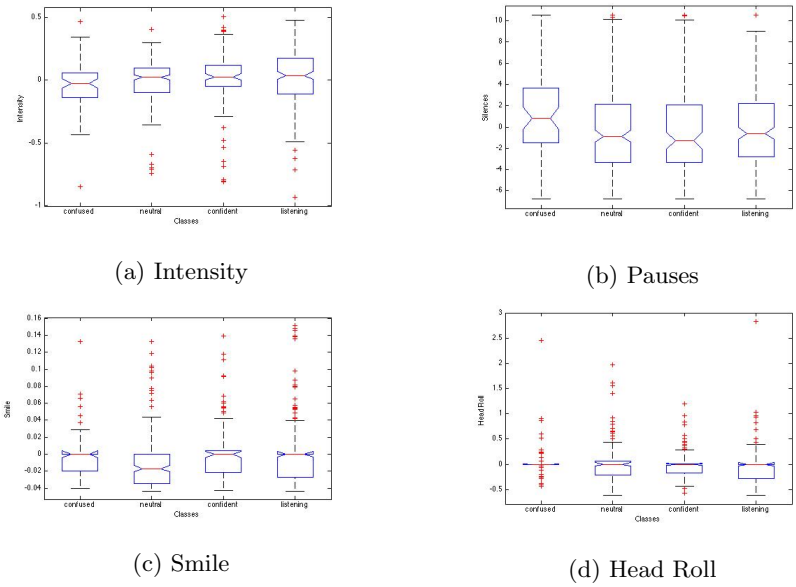
Fig. 2: Snapshots of frames showing the different affective states

The audio features were obtained from Praat[10] and they consisted of spectograms, pitch, formants, energy, and intensity patterns for each frame of the video. The video features were extracted from OKAO[8] and consisted of 165 features that included face pose, smile level, gaze direction, eye openness among many others. The box plots of some of the feature are as shown in Figs. 3. These suggested that the pauses and smiles are probably effective in classifying the affective states. The boxplots influenced our initial feature selection process.

5 Experimental Setup

The experimental setup can be divided into the following phases:

1. **Feature Extraction**: The feature extraction stage comprised of selecting the appropriate visual, audio cues and fusing them together using early fusion to form the multimodal dataset. The initial feature selection was done using the intuition derived from the qualitative analysis described in the previous section. For the audio feature set, energy, silence and intensity were chosen. Other features like formants and spectograms were not included as the dataset was quite noisy owing to the fact that there were two or more speakers at a time in certain frames, and it would be hard to filter out the noisy features in these cases. For the visual features, we chose the face pose (L/R and U/D), face directions (L/R, U/D and Roll) and the gaze directions (L/R, U/D), Eye Openness and Mouth Openness and Smile levels. The features for a particular frame were considered valid if and only if the face confidence was above a certain threshold. This was due to the fact that



(a) Intensity

(b) Pauses

(c) Smile

(d) Head Roll

Fig. 3: Boxplots for various features

there were multiple people in a few frames during transition of speakers, although there was best effort to clip the videos to contain a single face in each frame. Imposing confidence thresholds ensured that the intended speakers were indeed the subjects of the analysis. These features were further combined across frames to give a single value for each segment as follows:

- Energy: The standard deviation values of energy across frames was computed to give a single value for each segment that was annotated. This was done to normalize the difference in energy levels of voices from male and female participants.
- Intensity: The mean level of intensity was taken, as all students had a microphone and hence there was not an inherent need for normalization.
- Silence: Mean values were considered. This feature was later discarded as the variance was too less and hence proved to be useless in the task of classification.
- Face Pose: The standard deviation of the face pose was taken. This was due to the fact that the students were seated around a round table with the professor at the center. Thus, some faced left most times, and others right. The standard deviation accounted for feature normalization.
- Face Direction (up/down): The mean value was considered.
- Face Direction (left/right): Standard deviation values were considered, again to normalize the seating positions.
- Gaze Directions (up/down): Mean values were combined.
- Gaze Directions (left/right): Standard deviation values were considered for the same reasons listed previously.

– Eye Openness and Mouth Openness: Mean values were considered.

These features were further normalized with 0 mean and variance 1 to be able to work well with classifiers that assumed a Gaussian distribution of features.

2. **Feature Classification**: Linear Classifiers that have an independent assumption of features were considered as the segments of each speaker corresponded to a response to a single question, and hence were independent of each other. No student was ever asked two questions in sequence, the students always took turns to answer each question, and thus chances of the student showing the same emotion over the different segments was not very likely.

   We considered the Naive Bayes classifier and the Linear Support Vector Machines (SVM) classifier for the task. The entire dataset was split into 4 fold test and 3 fold validation sets to tune the parameters for the value of C in the case of the SVM classifier. The value of C corresponding to the best validation accuracy was then chosen to train the classifier and hence predict the test accuracy.

   All experiments were carried out in MATLAB using the MLToolbox provided by researchers at the Institute of Creative Technologies, affiliated with University of Southern California. The experiments involved training the classifiers with unimodal cues, and multimodal cues fused using early fusion techniques.

6 Results

In this section, we discuss the results of our experiments with the two different classifiers. Due to the heavy pre-processing and clipping that was required, our dataset consisted of videos where most of the segments were independent of each other. Hence, consideration of temporal dependencies of features would prove futile. Thus, the analysis was restricted to classifiers such as Naive Bayes and SVM that assume independence of features over segments.

6.1 Results with Naive Bayes Classifier

The Naive Bayes Classifer results for acoustic, visual and multimodal cues are as shown in Tables 1, 2 and 3 respectively. The validation accuracy (chosen as the performance measure) is in good agreement with the test accuracy, and this indicates that there is no overfitting involved. Furthermore, the Naive Bayes achieves a good accuracy of nearly 62% with the multimodal cues.

6.2 Results with Support Vector Machine Classifier

Support Vector Machine, known for better performance that Naive Bayes classifiers in a typical scenario, performed rather poorly in our task. The validation accuracy achieved was a maximum of 41.1% after 3-fold validation placing the

Table 1: Classification Results with Naive Bayes Classifier trained on Acoustic Features

	Train Accuracy	Validation Accuracy	Test Accuracy
Test Fold 1	0.5473	0.5285	0.5338
Test Fold 2	0.5473	0.5522	0.5338
Test Fold 3	0.5657	0.5244	0.5543
Test Fold 4	0.5473	**0.5548**	**0.6163**

Table 2: Classification Results with Naive Bayes Classifier trained on Visual Features

	Train Accuracy	Validation Accuracy	Test Accuracy
Test Fold 1	0.6483	0.6064	0.6110
Test Fold 2	0.6483	0.6060	0.6110
Test Fold 3	0.6285	0.5982	0.5888
Test Fold 4	0.6483	**0.6090**	**0.5338**

value of the cost parameter C at 10e2. The results with the SVM classifier are as shown in Table 4. The SVM trained on just one of the features did not yield any interesting results, and hence are not discussed here.

6.3 Possible Reason for Poor Performance of SVM:

We can see from 4 that SVM did not perform as well as the Naive Bayes. Our annotations were performed considering the baselines for each of the students. Hence, some students who speak very animatedly were marked "neutral" for the same cues for which others could have been marked as "confident/excited". This could have probably affected the accuracy of SVM. Increasing our soft margin parameter for SVM gave us slightly better results. This could be because higher tolerance to mislabeled data led SVM to better classify these labels which were associated with different cues in different data units.

6.4 Comparison of Unimodal and Multimodal Results

We have considered a random baseline. In this case, it was calculated by adding up the squared probabilities of each of the classes, which gave an accuracy of 30.016. The reason for the low accuracy is the uneven distribution of the classes (especially the confused class). We can see from Fig. 4 that all the different modalities perform better than the baseline. Furthermore, multimodal does perform the best among all the features, closely followed by visual features. One possible reason for this could be that as we are listening to another person, our first reaction to what is being said would result visually rather than verbally, in order to maintain etiquette and polite conversational rules.

In the audio cues, while cues such as voice energy and intensity were useful, they were very much dependent on the student's baseline. Visual cues on the

Table 3: Classification Results with Naive Bayes Classifier trained on Multimodal Features

	Train Accuracy	Validation Accuracy	Test Accuracy
Test Fold 1	0.6298	0.6117	0.6163
Test Fold 2	0.6554	**0.6137**	**0.5586**
Test Fold 3	0.6298	0.5758	0.6163
Test Fold 4	0.6298	0.5898	0.6110

Table 4: Classification Results with Support Vector Machine Classifier trained on Multimodal Features

	Train Accuracy	Validation Accuracy	Test Accuracy
Test Fold 1	0.5417	0.3423	0.3209
Test Fold 2	0.4871	**0.4111**	**0.4228**
Test Fold 3	0.4455	0.3556	0.3695
Test Fold 4	0.2249	0.2678	0. 2604

other hand were more general. For example, most students smiled when confused and raised their eyebrows while expressing strong views. This could also be a reason for the better performance of visual only modality as compared to audio only. Another important observation from the above graph is that the multimodal accuracy only increased by 0.5% from visual accuracy. This could be because of the fact that the number of audio features available was much fewer than that of visual.

6.5 Best Features for Classification

Video

1. Smile: From our observations of videos and statistical tests, we found Smile level as a very good indicator of confusion. From our statistical analysis, we noticed that smile levels are high for confusion.
2. Mouth Openness: This feature was very effective in differentiating listening video clips from speaking ones. The ANOVA test on this feature yielded a p-value of 0.0005, in agreement with the box plot (although not shown here) that produced a lower mean for mouth openness in the listener label as compared to all other labels.

Audio

1. Pauses: The students tend to have prolonged pauses when they are confused, as they tend to take time to think before they can put it into words. This, along with smile was the main indicator of the confused class. It was also observed that the duration and the number of pauses reduced as the confidence in the student increased.

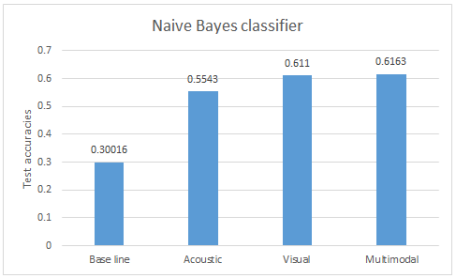


Fig. 4: Comparison of different modalities

2. Intensity: Intensity proved to be effective in not only identifying confident/excited speech, but also for differentiating the student listening videos (p=0.0532). Our box plots showed higher average intensities on videos where students were listening. This was probably because the professors voice had high intensity. The average intensities for the rest of the classes followed the expected trend - confused < neural < confident.

7 Conclusion and Future Work

From our experimental results, we have observed that Naive Bayes classifier outperformed the others. The reason for this we believe is that the independence of the features assumption might be in our favor in this case. Younger students tend to exhibit more prominent cues. Hence, our model might perform better when the subject is changed to a younger student. We can also see that the multimodal features definitely make the classifier better than each of the individual modalities.

7.1 Future Work

1. This model focuses mainly on classification while the student is speaking. It needs to be expanded to better classification while listening. This can be done with a larger dataset.
2. If clips pertaining to a single student are available, temporal data can be taken into consideration. In this way, the progress of the student can be measured as well. Classifiers such as Hidden Markov Model (HMM) and Hidden Conditional Random Field (HCRF) can be experimented with as well.
3. With the improvement in the tools available, more features can be considered such as body posture and hand gestures, which give a good insight of the students confidence level.
4. The focus in this study was only on the non verbal cues. Integrating verbal cues will give a much better performance when using the model with an ITS instead of a human tutor.

References

1. Grafsgaard, Joseph F., et al. The Additive Value of Multimodal Features for Predicting Engagement, Frustration, and Learning during Tutoring. Proceedings of the 16th International Conference on Multimodal Interaction. ACM, 2014.
2. Vail, Alexandria Katarina, and Kristy Elizabeth Boyer. Identifying Effective Moves in Tutoring: On the Refinement of Dialogue Act Annotation Schemes. Intelligent Tutoring Systems. Springer International Publishing, 2014.
3. Grafsgaard, Joseph F., et al. Automatically Recognizing Facial Expression: Predicting Engagement and Frustration. Proceedings of the 6th International Conference on Educational Data Mining. 2013.
4. D'Mello, Sidney K., and Arthur Graesser. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. User Modeling and User-Adapted Interaction 20.2 (2010): 147-187.
5. Sarrafzadeh, Abdolhossein, et al. How do you know that I don't understand? A look at the future of intelligent tutoring systems. Computers in Human Behavior 24.4 (2008): 1342-1363.
6. Modern & Contemporary American Poetry. `https://www.coursera.org/course/modernpoetry`
7. ELAN - The Language Archive. `http://tla.mpi.nl/tools/tla-tools/elan/`
8. OKAO Vision - Face Sensing Technology, `http://www.omron.com/r$_$d/coretech/vision/okao.html`
9. Framework for various Constrained Local Model based face tracking and landmark detection algorithms and their extensions/applications. `https://github.com/TadasBaltrusaitis/CLM-framework`
10. Doing phonetics using computer. `http://www.fon.hum.uva.nl/praat/`
11. A Cooperative Voice Analysis Repository for Speech Technologies. `https://github.com/covarep/covarep`