A Multimodal Framework for Automated Teaching Quality Assessment of One-to-many Online Instruction Videos

Yueran Pan\textsuperscript{ac}, Jiaxin Wu\textsuperscript{b}, Ran Ju\textsuperscript{c}, Ziang Zhou\textsuperscript{c}, Jiayue Gu\textsuperscript{b}, Songtian Zeng\textsuperscript{d}, Lynn Yuan\textsuperscript{d}, Ming Li\textsuperscript{ac}

\textsuperscript{a}School of Computer Science, Wuhan University, China
\textsuperscript{b}Center for Teaching and Learning, Duke Kunshan University, China
\textsuperscript{c}Data Science Research Center, Duke Kunshan University, China
\textsuperscript{d}Fumi Health & Technology LLC, China

ming.li369@dukekunshan.edu.cn

Abstract

In the post-pandemic era, online courses have been adopted universally. Manually assessing online course teaching quality requires significant time and professional pedagogy experience. To address this problem, we design an evaluation protocol and propose a multimodal machine learning framework\textsuperscript{1} for automated teaching quality assessment of one-to-many online instruction videos. Our framework evaluates online teaching quality from five aspects, namely Clarity, Classroom interaction, Technical management of online teaching, Empathy, and Time management. Our method includes mid-level behavior feature extraction, high-level interpretable feature extraction, and supervised learning prediction. Our automated multimodal teaching quality assessment system achieves comparable performance to human annotators on our one-to-many online instruction videos. For binary classification, the best average accuracy of five aspects is 0.898. For regression, the best average means square error is 0.527 on a 0-10 scale.

Keywords: Teaching Quality Assessment, Multimodal Behavior Coding, Interpretable Feature Extraction, Speaker Diarization, Emotion Recognition

I. Introduction

With the advent of the post-pandemic era, online courses have been adopted by an increasing number of institutions and service providers. With the screen, the teachers and students can utilize video conferences to realize real-time communication. However, it is hard for both sides to feel the other side’s real presence and have the same feeling as the in-person instruction sessions. Due to the sense of distance in online courses, the quality of classes may be affected to a certain extent. Because there are relatively few pedagogical experts, it is very hard for teachers to get timely feedback and adjust their teaching styles. It is also hard for all experts to reach the same benchmark and measurement when assessing videos just in an abstract sense. Hence, it is meaningful to propose an automated course teaching quality assessment system targeting online instruction to provide teachers with unified assessment standards and multi-dimensional quantized feedback that can be widely accepted and interpreted.

To explore this problem, we first collect a one-to-many online instruction video dataset and invite third-party education specialists to design an online teaching quality evaluation protocol for our dataset. Based on the evaluation metrics proposed by the educational specialists, we propose a multimodal machine learning framework to predict the teaching quality assessment scores.

In summary, the key contributions of this paper are:

\begin{itemize}
  \item We propose a two-level behavior feature extraction approach with both mid-level behavior descriptors and high-level clip-wise interpretable features to predict the online teaching quality assessment scores.
\end{itemize}

\textsuperscript{1} Framework details and demos are presented on https://github.com/sparklingyueran/Online-Course-Assessment-Framework
Speaker diarization, speech emotion recognition, automated speech recognition, facial expression recognition and head pose and gaze estimation are employed as the frontend mid-level behavior descriptors.

A fixed-dimensional, clip-wise and high-level interpretable feature vector is extracted on top of those mid-level behavior descriptors and serves as the input for the subsequent classifier.

II. Related Work and Motivation

For education evaluation, the transactional distance [1] can lead to some communication gaps, and it could hinder the teachers from understanding the real reactions of the students as well as prevent the students from conveying their feelings to the teachers in online teaching. According to [2], factors such as the class length, the contents of the instruction, and the interactions in class are highly associated with the quality of the online courses. Hone et al. [3] also found that interaction is crucial for the course evaluation as the activity connected both sides of the class. Moreover, a recent research [4], and the Moore’s model [5] both showed that a high-quality online teaching session would have good ratings in multiple dimensions such as level of difficulty, pressure, sense of belonging, time management, interactions, and engagement.


According to related works, there are many abstract concepts proposed to describe teaching quality in separate dimensions, and automated methods for score classification. There are two main limitations of recent studies: 1. Videos are only described in one dimension, and most time just labeled with a binary score. A widely accepted and multi-dimensional standard is lacking to describe and quantify teaching quality comprehensively. 2. Recent automated methods are not explainable enough, and teachers cannot understand the meaning of machine learning features. If more interpretable features are generated, it would be helpful for teachers to digest the feedback and improve their teaching quality.

In this case, we summarize previous pedagogical knowledge into a five-dimensional protocol and propose an explainable automated assessment framework on one-to-many lecture videos. We include the following five factors for evaluation: Clarity, Classroom interaction, Technical management of online teaching, Empathy, and Time management. Our model is specially designed for evaluating one-to-many adult online courses with one-to-one voice chat.

Clarity. Clarity is the quality of being coherent, logical and intelligible. It is considered to be what matters the most in teaching by Sharratt [11].

Classroom Interaction. In our case, it refers to the pedagogical activities focusing on the interaction between teachers and students [12], [13], such as the teacher’s instructional practices to invite students to interact with the teacher lively or via chat.

Technical Management of Online Teaching. The most essential difference between online and offline courses is that the carrier of online courses mainly depends on technology and the Internet. Thus, the technical management, which refers to how accurately and completely the information will be transmitted [14] including the teacher’s audio and video quality, the use of screen sharing and chat function of the platform is an indispensable factor.

Empathy. Empathy refers to the intention and ability to recognize, understand and resonate with the experience of the students [15]. Having empathy could help teachers provide instant support, adjust teaching contents flexibly and take advantage of the power of connection in teaching [16].

Time Management. In this fast-speed society, time management is an important ability, and it is also the case for online teaching. How to efficiently distribute the time of the essential class components is very important in online education.
III. Database Descriptio

Our database is collected by recording online lectures under the COVID-19 lockdown period with permission from all participants. Participants of each lecture include one teacher and 5-10 adult students. During the lecture, the teacher usually shares relevant slides and video demos on the screen. And sometimes, the teacher would invite the listeners to have an interactive activity such as having a voice chat. The face of the speaking listener would appear on the screen during the chat. Each lecture is delivered by a teacher who has received professional education about autism to several adult students who are autistic children’s parents. The content of the lectures is about child behavior development and autism behavioral intervention.

Our database contains 338 video clips from 63 lectures taught by 10 different teachers. All clips are split into sections based on the teaching purposes of that section. The types of the sections include experience sharing, opening & agenda, review of the previous study, main content, Q & A, scenarios & practices, intervention plan, a summary of the lecture, and future plan. The clip length varies from 5 to 30 minutes.

All clips are manually labeled according to the five dimensions: clarity, classroom interaction, technical management of online reaching, empathy, and time management. Three experienced pedagogy experts give a score independently from 0 to 10 in each dimension for every video clip. A higher rating represents better teaching quality. To be specific, 10 = Outstanding, 8= Excellent. 5=Good. 3= Fair. 0=Poor. Because online teachers in the real world can only work after training, most ratings are over 5, which is a qualified line for trained teachers. Only a few clips are rated lower than 5 in terms of classroom interaction because some sections (e.g., opening & agenda) do not require interaction. The rounded average number of all experts’ rates is set as the ground truth of each clip.

Each piece of data in our multimodal database contains the recorded video of that lecture, collected by DingTalk [17], an intelligent online communication platform, and recorded through the build-in local recording function. Our use of data is approved by the Institutional Review Board (IRB) of Duke Kunshan University. All participants have signed informed consent forms before lectures. Videos are recorded with permission from both the teachers and students. To protect participants’ privacy, all data are de-identified before the analysis, and the dataset is not open to access.

IV. Framework

Generally, our framework consists of three stages. First of all, we extract mid-level behavior features by multiple human behavior descriptors, including facial expression recognition, head pose and gaze estimation, voice activation detection, speaker diarization, and speech emotion recognition. Then we calculate 17 high-level interpretable statistical features on top of the outputs of mid-level modules to describe participants’ behavior. Finally, we combine all high-level features of a clip as a vector and adopt classification and regression methods to predict the teaching quality scores.

A. Mid-level Behavior Descriptors

1) Speaker Diarization: To partition audio into sentences and label each segment with teachers’ and students’ identities, we adopt a speaker diarization module. To better separate each speaker’s speech fragment and identify their speaking order, we choose a Bi-LSTM model [18] with spectral clustering. This vector-to-sequence model used a similarity matrix to compute scores between single speaker embedding and the whole embedding sequence and trained on CALLHOME database [19].

2) Speech Emotion Recognition: Since our task is language-specific in Chinese, we employed the Emotional Speech Dataset (ESD) to train the speech emotion recognizer [20], [21]. We utilized the librosa [22] library to extract logfbank feature from ESD. Next, we adopted ResNet-18 [23] to encode the audio features, which has been demonstrated effective in Speech Emotion Recognition (SER) tasks on multilingual datasets [24]. We performed the emotional classification by stacking several fully connected layers to the feature encoder.

3) Automated Speech Recognition: To obtain the transcript and further analyze the speech, we need to recognize the speech contents. Manual transcription demands enormous time and human efforts to complete and is therefore unrealistic. Alternatively, Automated Speech Recognition (ASR) offers a more
efficient and accurate option. We selected TDNN-F [25], the factorized form of TDN, to perform the ASR task, and pre-trained the model on a collection of AISHELL-2 corpus [26], MAGICDATA [27], and aidatatang_200zh [28].

FIG. 4: The framework of our facial expression recognition model.

TABLE I: The performance of different models on AffectNet dataset [30].

<table>
<thead>
<tr>
<th>Model</th>
<th>Valence Estimation(RMSE)</th>
<th>Arousal Estimation(RMSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mollahosseini et al [30]</td>
<td>0.397</td>
<td>0.142</td>
</tr>
<tr>
<td>Bulat et al [32]</td>
<td>0.356</td>
<td>0.326</td>
</tr>
<tr>
<td>Ours</td>
<td>0.3012</td>
<td>0.2078</td>
</tr>
</tbody>
</table>

4) Facial Expression Recognition: We utilize indicators of valence and arousal to describe the human’s facial emotion in a two-dimensional space instead of only adopting categorical facial expressions.

Based on Distract your Attention Network (DAN) [29], we modify the second half of the backbone model and add two additional prediction heads (MLP layers) to extend this classification-only model to regression of valence and arousal (Figure 4). Given that the network needs to take both classification (categorical facial expressions) and regression (valence and arousal) tasks, we train it on the database of AffectNet [30] which is a large-scale image-based emotion recognition database with the two target kinds of annotations. Moreover, we introduce the multi-task loss function [31] to balance the training of different task-related parts in the model. The designed loss function for multi-task optimization is described as follows:

\[
L = e^{-\sigma_1}L_{val} + e^{-\sigma_2}L_{af} + 2e^{-\sigma_3}(L_{cls} + L_{af}) + \sigma_1 + \sigma_2 + \sigma_3
\]

where \(\sigma_1\), \(\sigma_2\) and \(\sigma_3\) are the learnable parameters for the three outputs, \(L_{val}\) is the arousal loss, \(L_{af}\) is the valence loss, \(L_{cls}\) is the classification loss, \(L_{af}\) is the affinity loss, \(L_{pt}\) is the partition loss.

5) Head Pose and Gaze Estimation: As gaze movement performs an important role in human non-verbal cues, it is potential to consider introducing the gaze patterns to our framework. Thus, we implement the head pose and gaze estimation referred to Zhang’s method [33]. This model is composed of several convolutions with spatial attention-like architecture. Based on taking the full-face image as an appearance feature, it can predict the Euler Angle of the head pose and 3-D gaze direction. The model achieves the angular error of 4.8 on the MPIIGaze database [34] and 6.0 on the EYEDIAP [35] database.

B. Interpretable High-level Features

On top of the mid-level features extracted from the aforementioned speech processing and computer vision modules, we further design a set of interpretable high-level features for the whole clip.

1) Length of video is the time length of a video, including voice-active time and non-voice time. It can be directly extracted from the video file information. Because video lengths vary in the database, considering the length of the video can help avoid bias caused by it.

2) Speaker switch turns is the number of turn-talkings. The direct result of speaker diarization presents the speaker ID of each sentence. Only one person speaks at any point in videos because voice chats only happen between one adult student and one teacher. Counting how many times speaker ID changes in one video can get the number of turn-talkings.

3) Frequency of speaker switch is calculated by \(\frac{\text{Length of video}}{\text{Number of turn-talkings}}\), representing how frequently teachers invite students to have a communication or interaction.

4) Speaking duration ratio is calculated by \(\frac{\text{Speaker duration}}{\text{Length of video}}\). Speaker duration is the sum of voice-active time. The start time and the end time of each sentence are recorded in the speaker diarization outputs. \(\sum (\text{end time} - \text{start time})_{\text{each sentence}}\) is the speaker duration of each video. The speaking duration ratio is the active speaking proportion of the total time, reflecting whether the lecture is in an enthusiastic discussion or an awkward silence.

5) Average speech emotion is calculated by averaging the speech emotion recognition outputs along the time of each speaker. Each 5-second window is classified into a five-dimensional score vector, representing the probability of five discrete emotion categories, namely neutral, surprised, sad, happy, and angry. Both lectures’ and students’ average speech emotions are taken into consideration.

6) Percentage of speech emotion is the percentage of a certain speech emotion in the speech emotion classification time series. By choosing the emotion with the maximal probabilities in a vector, the speech emotion classification time series is calculated from the speech emotion probability vector sequence mentioned before. The percentage of a kind of speech emotion is calculated
V. Experiments

A. Experimental Settings

Voice activation detection: When we segment the audios, we use a 1.5s-duration and 750ms overlap sliding window to generate the homogeneous segments. And 750ms-long region in the center is the most talkative speaker for each segment.

Speech Emotion Recognition: When we extract the features, we choose window length as 0.025s, window step as 0.01s, the number of filters as 256, the FFT window size as 2048, the frequency range between 50 and 8000, and the pre-emphasis coefficient as 0.97. When training the classification model, we use SGD optimizer with the original learning rate as 0.01, momentum as 0.8, weight decay as 0.0001 and choose CosineAnnealingLR as the learning rate scheduler with 20 as the maximum number of iterations, $1e^{-8}$ as the minimum learning rate, setting verbose to True. Our classification accuracy reaches 0.8174.

Head Pose and Gaze Estimation: All the input images are regularized using the mean and variance values of ImageNet dataset [38]. When we train the net, we use Adam as the optimizer with the initial values of ImageNet dataset [38]. When we train the net, we use Adam as the optimizer with the initial learning rate as $10^{-4}$, momentum as 0.8, weight decay as 0.0001 and choose CosineAnnealingLR as the learning rate scheduler with 20 as the maximum number of iterations, $1e^{-8}$ as the minimum learning rate, setting verbose to True. Our classification accuracy reaches 0.8174.

B. Experimental Results

For binary classification (Table II), decision with random oversampling performs best. Generally, ran-
TABLE II: The accuracy for binary classification of different methods in five score dimensions. The score cutoff threshold is 8.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Clarity</th>
<th>Classroom interaction</th>
<th>Technical management</th>
<th>Empathy</th>
<th>Time management</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM + RO</td>
<td>0.740</td>
<td>0.849</td>
<td>0.740</td>
<td>0.740</td>
<td>0.704</td>
</tr>
<tr>
<td>SVM + SMOTE</td>
<td>0.728</td>
<td>0.860</td>
<td>0.762</td>
<td>0.740</td>
<td>0.794</td>
</tr>
<tr>
<td>LN + RO</td>
<td>0.732</td>
<td>0.778</td>
<td>0.698</td>
<td>0.698</td>
<td>0.690</td>
</tr>
<tr>
<td>LN + SMOTE</td>
<td>0.724</td>
<td>0.797</td>
<td>0.708</td>
<td>0.712</td>
<td>0.721</td>
</tr>
<tr>
<td>DT + RO</td>
<td>0.907</td>
<td>0.959</td>
<td>0.925</td>
<td>0.909</td>
<td>0.792</td>
</tr>
<tr>
<td>DT + SMOTE</td>
<td>0.835</td>
<td>0.890</td>
<td>0.840</td>
<td>0.802</td>
<td>0.724</td>
</tr>
</tbody>
</table>

TABLE III: Binary classification results of accuracy rate, precision rate, recall rate and F1 score by DT+RO.

<table>
<thead>
<tr>
<th>Score dimension</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarity</td>
<td>0.907</td>
<td>0.821</td>
<td>0.900</td>
<td>0.898</td>
</tr>
<tr>
<td>Classroom interaction</td>
<td>0.959</td>
<td>0.917</td>
<td>1.000</td>
<td>0.957</td>
</tr>
<tr>
<td>Technical management</td>
<td>0.925</td>
<td>0.849</td>
<td>1.000</td>
<td>0.918</td>
</tr>
<tr>
<td>Empathy</td>
<td>0.909</td>
<td>0.826</td>
<td>0.991</td>
<td>0.901</td>
</tr>
<tr>
<td>Time management</td>
<td>0.792</td>
<td>0.864</td>
<td>0.754</td>
<td>0.806</td>
</tr>
<tr>
<td>Mean</td>
<td>0.898</td>
<td>0.856</td>
<td>0.947</td>
<td>0.896</td>
</tr>
</tbody>
</table>

TABLE IV: MSE results against the ground truth of five dimensions for pedagogy experts and different regression methods.

<table>
<thead>
<tr>
<th>Scoring dimensions</th>
<th>Clarity</th>
<th>Classroom interaction</th>
<th>Technical management</th>
<th>Empathy</th>
<th>Time management</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert 1</td>
<td>0.869</td>
<td>2.902</td>
<td>1.380</td>
<td>1.364</td>
<td>1.479</td>
</tr>
<tr>
<td>Expert 2</td>
<td>1.000</td>
<td>2.306</td>
<td>0.923</td>
<td>1.346</td>
<td>1.479</td>
</tr>
<tr>
<td>Expert 3</td>
<td>0.380</td>
<td>4.264</td>
<td>1.644</td>
<td>1.669</td>
<td>0.850</td>
</tr>
<tr>
<td>Linear regression</td>
<td>0.277</td>
<td>4.112</td>
<td>0.410</td>
<td>0.498</td>
<td>0.283</td>
</tr>
<tr>
<td>KNN</td>
<td>0.262</td>
<td>4.449</td>
<td>0.451</td>
<td>0.422</td>
<td>0.292</td>
</tr>
<tr>
<td>SVM</td>
<td>0.235</td>
<td>4.060</td>
<td>0.383</td>
<td>0.413</td>
<td>0.278</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>0.257</td>
<td>4.345</td>
<td>0.286</td>
<td>0.374</td>
<td>0.275</td>
</tr>
<tr>
<td>RandomForest</td>
<td>0.249</td>
<td>1.381</td>
<td>0.384</td>
<td>0.365</td>
<td>0.291</td>
</tr>
</tbody>
</table>

TABLE V: The average MSE, MAE and PPMCC of five rating dimensions against the ground truth for pedagogy experts and regression of different methods.

<table>
<thead>
<tr>
<th>Results</th>
<th>MSE</th>
<th>MAE</th>
<th>PPMCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert 1</td>
<td>1.782</td>
<td>0.970</td>
<td>0.503</td>
</tr>
<tr>
<td>Expert 2</td>
<td>1.410</td>
<td>0.910</td>
<td>0.684</td>
</tr>
<tr>
<td>Expert 3</td>
<td>1.761</td>
<td>0.981</td>
<td>0.509</td>
</tr>
<tr>
<td>Linear regression</td>
<td>0.576</td>
<td>0.564</td>
<td>0.423</td>
</tr>
<tr>
<td>KNN</td>
<td>0.575</td>
<td>0.548</td>
<td>0.457</td>
</tr>
<tr>
<td>SVM</td>
<td>0.541</td>
<td>0.513</td>
<td>0.501</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>0.527</td>
<td>0.549</td>
<td>0.507</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.528</td>
<td>0.526</td>
<td>0.497</td>
</tr>
</tbody>
</table>

TABLE VI: The scoring dimensions in the pedagogical model with their corresponding interpretable high-level features. (Per: percentage of, AV: Average of, AVN: Average number of, F: Frequency of, TTR: Type token ratio, NDW: Number of different words, SE: Speech emotion, SS: speaker switch, SPS: Speed of sentences, SDR: Speaking duration ratio.)

VI. Conclusion and Future Works

In this paper, we propose a multimodal framework for automated teaching quality assessment, integrating speaker diarization, facial emotion recognition, and other speech and computer vision machine learning methods. Our framework could fit the professional manual judgment well. For regression, the best average MSE is 0.527, and our PPMCC results are close to experts’. It shows that its prediction is comparable with single experts’ annotation. Compared to manual rating, it has a more clear and stable benchmark for assessment than the human sense, which could effectively reduce errors and fill the need for rating specialists. For binary classification, the best average accuracy of five aspects is 0.898, and ‘not excellent’ lectures can be selected efficiently. Related to-be-improved features can be used for the follow-up education training. To the best of our knowledge, it is the first system integrating speaker-diarization-based features to assess the online course. In this case, it would contribute to improving online teaching quality.

Considering assessment is a very subjective work, we admit that referring to three experts’ scores as the ground truth may not be sufficient.

In future work, we plan to collaborate with more education experts and test the compatibility between our framework and lecture for other topics. We plan to propose an automated framework that works for more kinds of education.

Acknowledgment

This research is funded in part by the National Natural Science Foundation of China (62171207), Science and Technology Program of Guangzhou City (202007030011), DKU Interdisciplinary Seed Grant and Dami&Xiaomi.
References


