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

Yueran Pan[ac], Jiaxin Wu[b], Ran Ju[c], Ziang Zhou[c], Jiayue Gu[b], Songtian Zeng[d], Lynn Yuan[d], Ming Li[ac]
[a]*School of Computer Science, Wuhan University, China*
[b]*Center for Teaching and Learning, Duke Kunshan Univerisity, China*
[c]*Data Science Research Center, Duke Kunshan University, China*
[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[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, Interptretable 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.

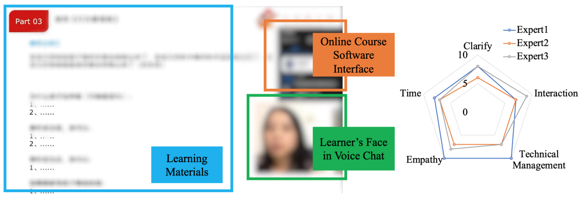

Fig. 1: LEFT: An example of the online course video. The main part of the shared screen presents slides or other kinds of learning materials. The top right corner shows the interface of the online course software. The face of the student who speaks in the voice chat would show in the bottom right corner. RIGHT: A radar chart presents scores from three experts for a clip.

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:

- 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.

[1] Framework details and demos are presented on https://github.com/sparklingyueran/Online-Course-Assessment-Framework

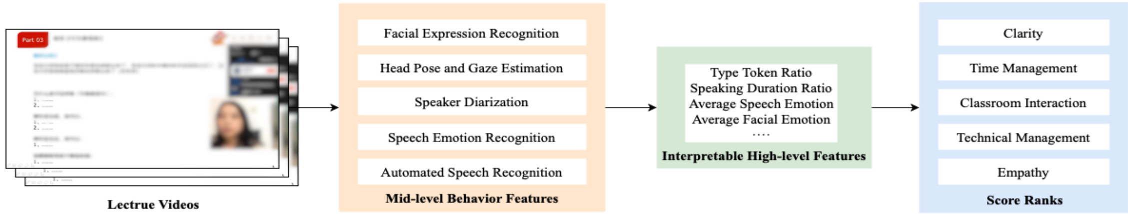

Fig. 2: Our automated framework for online course teaching quality assessment.

- 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.

For automated assessment, many studies apply machine learning to audio and video data to improve assessment efficiency. Zualkernan[6] developed an audio-based CNN for class observation classification. Li[7] applied multimodal CNN on data of facial attentive expression, speech text, and heart rate to infer student's attention to help reflect teaching quality. According to the CLASS[8] observation protocol, Anusha[9] and Ramakrishnan[10] proposed a multimodal framework to do binary classification of positive/negative climate.

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

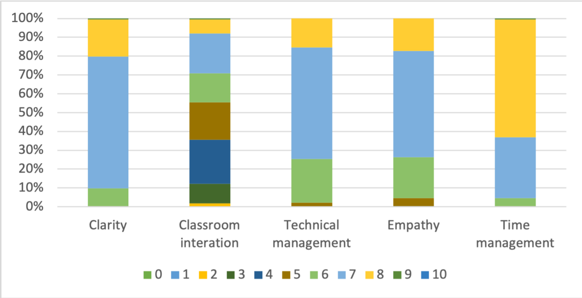

Fig. 3: Score distribution of our database.

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 participators' 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

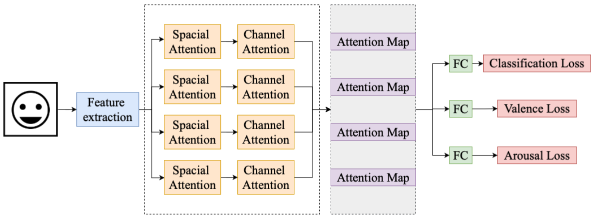

Fig. 4: The framework of our facial expression recognition model.

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

Model	Valence Estimation(RMSE)	Arousal Estimation(RMSE)
Mollahosseini et.al [30]	0.360	0.342
Bulat et.al [32]	0.356	0.326
Ours	**0.3012**	**0.2978**

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].

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:

$$L = e^{-\sigma_1} L_{as} + e^{-\sigma_2} L_{vc} + 2e^{-\sigma_3}(L_{cls} + L_{af} + L_{pt}) + \sigma_1 + \sigma_2 + \sigma_3$$

where σ_1, σ_2 and σ_3 are the learnable parameters for the three outputs, L_{as} is the arousal loss, L_{vc} is the valence loss, L_{cls} is the classification loss, L_{af} is the affinity loss, L_{pt} is the partion 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 $\frac{Speaker\ switch\ turns}{Length\ of\ video}$, representing how frequently teachers invite students to have a communication or interaction.
4) **Speaking duration ratio** is calculated by $\frac{Speaker\ duration}{Length\ of\ video}$. Speaker duration is the sum of voice-active time. The start time and the end time of each sentence are recorded in the speaker diarizatoin outputs. $\sum(end\ time - start\ time)_{each\ sentence}$ 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

by $\frac{Number\ of\ the\ emotion}{Length\ of\ the\ emotion\ classification\ sequence}$. This feature is used to describe teachers' emotions and students' emotions generally.

7) **Total number of words (TNW)** counts all teachers' words in a clip [36]. From automated speech recognition, teachers' audios are transferred as words without punctuation. The total number of words represents how much content a teacher conveys in a video.
8) **Number of different words (NDW)** counts the number of unique words in a video clip, reflecting teachers' vocabulary diversity [36].
9) **Type token ratio (TTR)** is calculated by NDM/TNW. Type token ratio shows how frequently teachers adopt new vocabulary. It indicates a teacher's word variety.
10) **Average number of different words** is calculated by $NDW/(Number of sentences)$. It is for measuring a teacher's language variety per sentence.
11) **Average speed of sentences** is the mean of speaking speed for each sentence, representing how quickly a teacher delivers sentences averagely.
12) **MLUw** is mean length of utterances between words [37]. It describes the average duration of a teacher's sentence. It generally measures a teacher's ability in language organization.
13) **MLU5w** is the mean length of words of each speaker's five longest utterances [37]. It represents a teacher's best performance in language organization.
14) **Average head pose and gaze** is the mean of students' head and gaze angle when in voice chat. It represents whether a student focuses on the screen generally.
15) **Average facial emotion** is calculated by averaging possible time series for each facial emotion class. The calculation is similar to the mean of speech emotion. For facial emotion, we have seven classes of emotion: neutral, anger, happy, sad, fear, surprised, and disgusted.
16) **Average arousal** is calculated by averaging the arousal time series for each facial emotion, representing the degree of emotion activation. It measures how much a student's emotion is activated by a teacher.
17) **Average valence** is calculated by averaging the valence time series for each facial emotion, representing the degree of pleasant emotion. It measures how pleasant a student is when chatting with a teacher.

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 learning rate as 10^{-4}, β_1 as 0.9 and β_2 as 0.999.

All interpretable high-level features are min-max normalized before the final backend predictor of the framework. We perform both binary classification and regression to predict scores. For binary classification, We set rating 8 as the cutoff threshold and merge rating below 8 because (1)the scoring distribution mainly lies between 5 to 10; (2)filtering 'excellent' teachers is instructive for teaching quality evaluation. We employ SVM, logistic regression (LN) and decision tree (DT) for comparison. Because the amount of the database is limited, we apply leave-one-out cross-validation (LOOCV) to enhance model usability, that in each cross, we leave one lecture out. Random oversampling (RO) and Synthetic Minority Oversampling Technique (SMOTE) is performed to reduce the possible impact of imbalance distribution. For regression, we employ linear regression, KNN, SVM, AdaBoost and random forest. We compare regression results with every single pedagogical expert's result to evaluate the efficiency of our framework. We adopt Mean Absolute Error (MAE), Mean Squared Error (MSE) and Pearson product-moment correlation coefficient (PPMCC) to evaluate methods.

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 dimension. The score cutoff threshold is 8.

Methods	Clarity	Classroom interaction	Technical management	Empathy	Time management
SVM + RO	0.740	0.847	0.740	0.740	0.704
SVM + SMOTE	0.728	0.869	0.762	0.740	0.709
LN + RO	0.732	0.778	0.698	0.698	0.699
LN + SMOTE	0.724	0.797	0.708	0.717	0.721
DT + RO	**0.907**	**0.959**	**0.925**	**0.909**	**0.792**
DT + SMOTE	0.835	0.890	0.840	0.802	0.724

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

Score dimension	Accuracy	Precision	Recall	F1
Clarity	0.907	0.821	0.990	0.898
Classroom interaction	0.959	0.917	1.000	0.957
Technical management	0.925	0.849	1.000	0.918
Empathy	0.909	0.826	0.991	0.901
Time management	0.792	0.864	0.754	0.806
Mean	0.898	0.856	0.947	0.896

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

Score dimension	Clarity	Classroom interaction	Technical management	Empathy	Time management
Expert 1	0.629	4.902	1.356	1.187	0.837
Expert 2	1.000	2.3067	0.923	1.344	1.479
Expert 3	0.380	4.264	1.644	1.669	0.850
Linear regression	0.277	1.412	0.410	0.498	0.283
KNN	0.262	1.449	0.451	0.422	0.292
SVM	**0.235**	1.400	0.383	0.413	**0.273**
AdaBoost	0.257	1.345	**0.382**	**0.374**	0.275
RandomForest	0.2469	**1.331**	0.3854	0.3865	0.291

dom oversampling methods improve the problem of the imbalanced dataset better than SMOTE methods. Especially, decision tree with random oversampling (Table III) reaches the best results for all rating dimensions, that most are over 0.9. The best accuracy is 0.959 for classroom interaction. The average accuracy by decision with random oversampling of five dimensions is 0.898.

The purpose of employing binary classification is to help filter 'excellent' lectures and 'good' lectures. In this case, the recall rate of 'not excellent' lectures is meaningful for advice. The average recall rate of five score dimensions reaches 0.9471, which means that our framework can significantly help filter out lectures that remain to be improved.

For regression (Table IV V), all our proposed methods perform better than single experts' result for MSE, MAE and PPMCC. Especially, SVM, random forest and AdaBoost perform well. AdaBoost reaches the best average MSE at 0.527 of five rating dimensions. Our PPMCC results are close to expert 1's and expert 3's, while expert 2's PMCC is the highest one. It illustrates that our proposed method can provide comparable performance with manual rating.

VI. Conclusion and Future Works

In this paper, we propose a multimodal framework for automated teaching quality assessment, integrating speaker diarizatoin, facial emotion recognition, and

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

Results	MSE	MAE	PPMCC
Expert 1	1.782	0.970	0.503
Expert 2	1.410	0.910	**0.684**
Expert 3	1.761	0.981	0.509
Linear regression	0.576	0.564	0.423
KNN	0.575	0.548	0.457
SVM	0.541	0.513	0.501
AdaBoost	**0.527**	0.549	0.507
Random Forest	0.528	**0.526**	0.497

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.)

Scoring dimensions	Corresponding Interpretable High-level Features				
Clarity	Per SE	AV SE	MLUw	MLU5w	AV SPS
Classroom interaction	F SS	AV SE	SS turns	SDR	TTR
Technical management	Per SE	AV SE	MLUw	MLU5w	AV NDW
Empathy	Per SE	AV SE	MLUw	MLU5w	AV SPS
Time management	NDW	AV SE	MLUw	Length of Video	TTR

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

[1] M. G. Moore, "The theory of transactional distance," in *Handbook of distance education*. Routledge, 2013, pp. 84–103.

[2] K. Jordan, "Massive open online course completion rates revisited: Assessment, length and attrition," *International Review of Research in Open and Distributed Learning*, vol. 16, no. 3, pp. 341–358, 2015.

[3] K. S. Hone and G. R. El Said, "Exploring the factors affecting mooc retention: A survey study," *Computers & Education*, vol. 98, pp. 157–168, 2016.

[4] R. Deng, P. Benckendorff, and D. Gannaway, "Progress and new directions for teaching and learning in moocs," *Computers & Education*, vol. 129, pp. 48–60, 2019.

[5] G. Falloon, "Making the connection: Moore's theory of transactional distance and its relevance to the use of a virtual classroom in postgraduate online teacher education," *Journal of Research on Technology in Education*, vol. 43, no. 3, pp. 187–209, 2011.

[6] I. Zualkernan and M. S. Khan, "Towards an audio-based cnn for classroom observation on a smartwatch," in *2020 IEEE/ITU International Conference on Artificial Intelligence for Good (AI4G)*. IEEE, 2020, pp. 224–229.

[7] J. Li, R. Li, Y. Zhou, J. Xian, X. Zhang, and X. Hei, "Inferring student's attention in a machine learning approach: A feasibility study," in *2019 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*. IEEE, 2019, pp. 1–2.

[8] R. C. Pianta, K. M. La Paro, and B. K. Hamre, *Classroom Assessment Scoring System™: Manual K-3*. Paul H Brookes Publishing, 2008.

[9] A. James, M. Kashyap, Y. H. V. Chua, T. Maszczyk, A. M. Núñez, R. Bull, and J. Dauwels, "Inferring the climate in classrooms from audio and video recordings: a machine learning approach," in *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*. IEEE, 2018, pp. 983–988.

[10] A. Ramakrishnan, B. Zylich, E. Ottmar, J. LoCasale-Crouch, and J. Whitehill, "Toward automated classroom observation: multimodal machine learning to estimate class positive climate and negative climate," *IEEE Transactions on Affective Computing*, 2021.

[11] L. Sharratt, *Clarity: What matters most in learning, teaching, and leading*. Corwin Press, 2018.

[12] H. Z. Waring, "Moving out of irf (initiation-response-feedback): A single case analysis," *Language Learning*, vol. 59, no. 4, pp. 796–824, 2009.

[13] Y. Matsumoto and A. M. Dobs, "Pedagogical gestures as interactional resources for teaching and learning tense and aspect in the esl grammar classroom," *Language Learning*, vol. 67, no. 1, pp. 7–42, 2017.

[14] C. E. Shannon and W. Weaver, "A mathematical model of communication," *Urbana, IL: University of Illinois Press*, vol. 11, 1949.

[15] J. V. Jordan, "Relational-cultural therapy," *Handbook of counseling women*, pp. 63–73, 2010.

[16] J. V. Jordan and H. L. Schwartz, "Radical empathy in teaching," *New Directions for Teaching and Learning*, vol. 2018, no. 153, pp. 25–35, 2018.

[17] A. Group, "Dingtalk," https://www.dingtalk.com/en.

[18] Q. Lin, R. Yin, M. Li, H. Bredin, and C. Barras, "LSTM Based Similarity Measurement with Spectral Clustering for Speaker Diarization," in *Proceedings of Interspeech*, 2019, pp. 366–370.

[19] W. Wang, D. Cai, X. Qin, and M. Li, "The dku-dukeece systems for voxceleb speaker recognition challenge 2020," *arXiv preprint arXiv:2010.12731*, 2020.

[20] "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[21] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 920–924.

[22] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[24] Z. Zhou, Y. Xu, and M. Li, "Detecting escalation level from speech with transfer learning and acoustic-lexical information fusion," 2021.

[25] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in *Interspeech*, 2018, pp. 3743–3747.

[26] J. Du, X. Na, X. Liu, and H. Bu, "Aishell-2: Transforming mandarin asr research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.

[27] L. Magic Data Technology Co., "Magicdata mandarin chinese read speech corpus," http://www.imagicdatatech.com/index.php/home/dataopensource/data_info/id/101",05/2019, 2019.

[28] L. Beijing DataTang Technology Co., "aidatatang_200zh, a free chinese mandarin speech corpus," www.datatang.com.

[29] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: Multi-head cross attention network for facial expression recognition," 2021.

[30] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[31] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of IEEE conference on computer vision and pattern recognition*, 2018, pp. 7482–7491.

[32] A. Bulat, S. Cheng, J. Yang, A. Garbett, E. Sanchez, and G. Tzimiropoulos, "Pre-training strategies and datasets for facial representation learning," *arXiv preprint arXiv:2103.16554*, 2021.

[33] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 51–60.

[34] ——, "Appearance-based gaze estimation in the wild," in *Proceedings of IEEE conference on computer vision and pattern recognition*, 2015, pp. 4511–4520.

[35] K. A. Funes Mora, F. Monay, and J.-M. Odobez, "Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, 2014, pp. 255–258.

[36] M. J. Pezold, C. M. Imgrund, and H. L. Storkel, "Using computer programs for language sample analysis," *Language, Speech, and Hearing Services in Schools*, vol. 51, no. 1, pp. 103–114, 2020.

[37] B. MacWhinney, *The CHILDES Project: Tools for analyzing talk. transcription format and programs*. Psychology Press, 2000, vol. 1.

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.