Understanding the Teaching Styles by an Attention based Multi-task Cross-media Dimensional modelling

Suping Zhou
Department of Computer Science and Technology
Tsinghua University
1874504489@qq.com

Jia Jia*
Tsinghua University
Beijing National Research Center for Information Science and Technology
jjia@mail.tsinghua.edu.cn

Yufeng Yin
Tsinghua University
yinyf15@mails.tsinghua.edu.cn

Xiang Li
Tsinghua University
li-xiang16@mails.tsinghua.edu.cn

Yang Yao
Tsinghua University
yanghaiqie@outlook.com

Ying Zhang
Tsinghua University
zhangy0018@163.com

Zeyang Ye
Tsinghua University
yezeyang17@gmail.com

Kehua Lei
Tsinghua University
492808177@qq.com

Yan Huang
TAL Education Group
galehuang@100tal.com

Jialie Shen
Queen's University of Belfast
j.shen@qub.ac.uk

ABSTRACT

Teaching style plays an influential role in helping students to achieve academic success. In this paper, we explore a new problem of effectively understanding teachers' teaching styles. Specifically, we study 1) how to quantitatively characterize various teachers' teaching styles for various teachers and 2) how to model the subtle relationship between cross-media teaching related data (speech, facial expressions and body motions, content et al.) and teaching styles. Using the adjectives selected from more than 10,000 feedback questionnaires provided by an educational enterprise, a novel concept called Teaching Style Semantic Space (TSSS) is developed based on the pleasure-arousal dimensional theory to describe teaching styles quantitatively and comprehensively. Then a multi-task deep learning based model, Attention-based Multi-path Multi-task Deep Neural Network (AMMDNN), is proposed to accurately and robustly capture the internal correlations between cross-media features and TSSS. Based on the benchmark dataset, we further develop a comprehensive data set including 4,541 full-annotated cross-modality teaching classes. Our experimental results demonstrate that the proposed AMMDNN outperforms (+0.0842 in terms of the concordance correlation coefficient (CCC) on average) baseline methods. To further demonstrate the advantages of the proposed TSSS and our model, several interesting case studies are carried out, such as teaching styles comparison among different teachers and courses, and leveraging the proposed method for teaching quality analysis.

*corresponding author.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3351059

CCS CONCEPTS

• **Social and professional topics** → **Computing education**; • **Applied computing** → **Education**; • **Computing methodologies** → *Multi-task learning*; *Neural networks*.

KEYWORDS

Teaching styles, Multi-task, Attention

ACM Reference Format:
Suping Zhou, Jia Jia, Yufeng Yin, Xiang Li, Yang Yao, Ying Zhang, Zeyang Ye, Kehua Lei, Yan Huang, and Jialie Shen. 2019. Understanding the Teaching Styles by an Attention based Multi-task Cross-media Dimensional modelling. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3343031.3351059

1 INTRODUCTION

Teaching effectiveness has a significant influence on student's learning performance [1]. In real practice, teachers assist students to achieve learning objectives mainly through verbal and nonverbal behaviours with multiple channels [9]. Generally, the teaching behavior and the teaching beliefs matching to it can be defined as *teaching style*[11][10], which can be conveyed by teachers' speech, body motion, as well as teaching contents during the class. Consequently, this inspires us about potential of leveraging cross-media data to model teachers' verbal and nonverbal behaviours and gain comprehensive understanding of teaching styles.

Teaching is complex process and thus how to develop advanced technique to characterize teaching styles based on multi-modal learning is not a trivial task. Several related attempts have been witnessed. Zhou *et al.* [28] Propose a Multi-path Generative Neural Network which considers both acoustic and textual features for emotion inferring. Chen *et al.* [5] Propose a novel scheme for Twitter sentiment analysis with textual information and extra attention to emojis. Focus on style analysis, Kwon *et al.* [13] uses Gaussian SVM to conduct a four-class speaking styles classification of three stressed styles (angry, Lombard and loud) and a neutral style. Mohammadi *et al.* [18] applies prosodic features and personality assessments to distinguish between professional and non-professional speaking styles. However, works about teaching styles inferring are limited, and most of them are single modal analysis [13][21].

In this paper, we aim to leverage cross-media information about teaching classes including acoustic, visual and textual data to achieve effective teaching style understanding. Towards this end, we address two important questions: 1) how to quantitatively characterize various teachers' teaching styles for various teachers, 2) how to model the subtle relationship between cross-media teaching related data and teaching styles. In order to solve the above challenges, firstïijŇa two-dimensional Teaching Style Semantic Space (TSSS) is built to describe teaching styles quantitatively and comprehensively based on the pleasure-arousal dimensional theory proposed by [16]. Then based on Thulac [23], the most often used 41 teaching style adjectives are selected from more than 10,000 feedback questionnaires provided by Tomorrow Advancing Life (TAL) Education Group [1], and manually labeled them on the TSSS. Also an Attention based Multi-path Multi-task Deep Neural Network (AMMDNN) is proposed to accurately and robustly capture the internal correlations between cross-media features and TSSS, which consists of pleasure-arousal values, and adjective words. Employing the benchmark dataset we build with 4,541 cross-media teaching classes data collected from TAL, an extensive range of tests have been designed to evaluate the mapping effects between cross-media features and coordinate values in the TSSS. The results indicate that the proposed AMMDNN model outperforms (+0.0842 in terms of CCC on average) several alternative baselines. Meanwhile, by linking the two-dimensional coordinates with teaching style adjectives, we further show that our method can help to describe the teaching styles more reasonably and vividly. Finally, we also carry out several interesting case studies including teaching styles comparison among different teachers and courses, and leveraging the proposed method for teaching quality analysis. Given the importance of teaching style, this study can serve as a springboard for further scholarly exploration.

Our contributions can be summarized as follows:

- We explore a new problem of effectively understanding teachers' teaching styles based on multi-modal learning. To our best knowledge, this is the first attempt to study the problem of understanding teachers' teaching styles automatically. Moreover, some interesting case studies are carried to show the effectiveness of our proposed method for understanding teaching styles, such as teaching styles comparison among different teachers and courses, teaching quality analysis.
- To gain a comprehensive and flexible teaching styles description, we adopt a dimensional method, a universal Teaching Style Semantic Space (TSSS) to describe teaching styles quantitatively. It is a two-dimensional space containing 41 words that people often use to describe teaching styles. Based on the TSSS, we can not only quantitatively analyze the teachers' teaching styles but also describe the teaching styles more reasonably and vividly.
- We propose an Attention based Multi-path Multi-task Deep Neural Network (AMMDNN), to implement the task of mapping cross-media features of teachers in class to the TSSS. Specifically, first we use a multi-path solution to avoid high dimensional input with limit training data, then we adopt attention mechanism to merge the high-level representations of multi-modal features which can better capture the cross-modality correlations in element samples. Furthermore, we regard predicting pleasure and arousal values as two related tasks to further improve the performance.

[1]http://en.100tal.com/ a leading education and technology enterprise in China

The rest of paper is organized as follows. Section 2 lists related works. Section 3 formulates the problem. Section 4 presents the methodologies. Section 5 introduces the dataset, experiment results and case studies. Section 6 is the conclusion.

2 RELATED WORK

Analysis on personal styles. Many efforts have been made on personal styles analysis in different application domain [20]. In [26], Wurtzel *et al.* state that acting styles for television and cinema tend to be controlled and naturalistic, while stage acting style is expansive and somewhat exaggerated. Smiljanic *et al.* [21] develop a method to elicit three different speaking styles: reduced, citation, and hyper-articulated. Using isolated words recorded at an 8kHz sampling rate in various speaking styles, Kwon *et al.* [13] define three stressed styles (angry, Lombard and loud) and one neutral style, and use Gaussian SVM to conduct four-class speaking style classification. Eyben *et al.* [7] utilize SVM to classify different dance styles like Waltz, Viennese Waltz, Tango, Quick Step, Foxtrot, Rumba, Cha Cha, Samba and so on. However, these studies mainly focus the categorical styles analysis, which may limit the diversity of personal styles.

Cross-media dimensional modelling. There are two main types of cross-media modelling strategies: categorical and dimensional ones. For categorical modelling, Chen *et al.* [5] propose a novel scheme for Twitter sentiment analysis with textual information and extra attention on emojis. Zhou *et al.* [28] propose a Multi-path Generative Neural Network which considers both acoustic and textual features for emotion inferring. Since categorical modelling may limit the diversity of personal styles, plenty of previous works based on dimensional modelling have been done in recent years. Thammasan *et al.* [24] present a multi-modal study of fusion of EEG and musical features in recognition of arousal and pleasure values for music. Chen *et al.* [4] apply a multi-task learning strategy for multiple kinds of para-linguistic information with shared representations.

Nevertheless, works about how to build a semantic space for teaching styles analysis are limited, and most of them are based on single modality analysis. Therefore, how to map cross-media information in the teaching classes to the teaching styles accurately is still a problem.

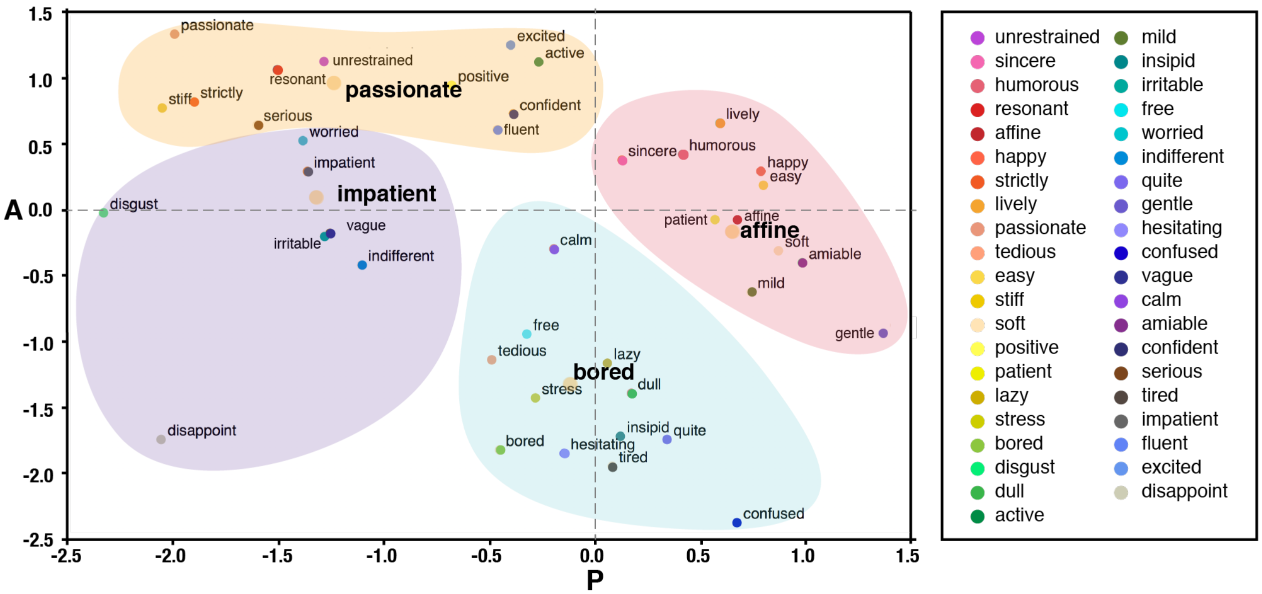

Figure 1: The Teaching Style Semantic Space

3 PROBLEM FORMULATION

Given a set of utterances V, for each utterance $v \in V$, we denote $v = \{x^a, x^t, x^v\}$. x^a represents the acoustic features of each utterance, which is a n_a dimensional vector. x^t represents the textual features of each utterance, which is a n_t dimensional vector.x^v represents the visual features of each utterance, which is a n_v dimensional vector. In addition, X^a is defined as a $|V| * n_a$ feature matrix with each element x^a_{ij} denoting the jth acoustic feature of v_i. The definition of X^t and X^v is similar to X^a.

Definition **The teaching style semantic space** - We adopt a two-dimensional space (pleasure and arousal), denoted as $D(p, a)$. The horizontal axis represents coordinate value for pleasure, while the vertical axis represents coordinate value for arousal.

Problem **Learning task** - Given utterances set V, we aim to infer the coordinate value in the Teaching Style Semantic Space for every utterance $v \in V$:

$$f : (V, X^a, X^t, X^v) \Rightarrow D(p, a) \tag{1}$$

4 METHODOLOGY

In this study, teaching styles understanding can be formulated with three subtasks: 1) Building a Teaching Style Semantic Space (TSSS) to describe teaching styles systematically and quantitatively; 2) Utilizing different regression models to build the mapping from cross-media features to the TSSS; 3) Linking the two-dimensional coordinates with teaching style adjectives.

4.1 Building the Teaching Style Semantic Space

Inspired by [17], a two-dimensional semantic space is proposed to describe various teachers' teaching styles quantitatively. Specifically, the semantic space has two dimensions. The horizontal axis refers to pleasure value indicating how pleasant or unpleasant teachers feel, while the vertical axis refers to arousal value indicating how energized or soporific teachers feel. However, two questions need to be addressed for building the TSSS: 1) What adjectives should we use in the semantic space? 2) How can we map the teaching style adjectives to the semantic space?

To describe different teachers' teaching styles, we first observe the feedback questionnaires from parents to evaluate the teachers. The questionnaires are from TAL Education Group which contains over 10,000 surveys. And then, the adjectives about teaching styles are selected based on a Chinese text processing tool Thulac [23]. As a result, 505 words are obtained, and through the manually selection, we remove those adjectives that not often used to describe teaching styles and finally gain the word list containing 41 teaching style adjectives. Furthermore, to have a more concise and clear description of teaching style, we separate 41 teaching style adjectives into four major teaching style categories, where the teaching styles in each major category have similar verbal and nonverbal behaviours. The details of the 41 teaching style adjectives and four major categories are shown in Figure 2.

Then we need to assign two-dimensional coordinates for each adjective. First, five different annotators are also invited to label the utterances with multiple teaching style adjectives. When more than three annotators choose an adjective for one utterance, this utterance is mapped to this adjective. Finally, for one teaching style adjective, it has a set of utterances mapping to it. The labelling methods of pleasure and arousal values for each utterance are presented in Section 5.2 and Table 3. As we see, the center point of all the utterances is mapped to the teaching style adjective as its pleasure and arousal value.

With this way, we can build the Teaching Style Semantic Space shown in Figure 1. Specifically, the horizontal axis refers to pleasure value while the vertical axis refers to arousal value. Each point in the figure represents a teaching style adjective. We can see those positive adjectives such as humor, lively and friendly lay in the top

Passionate		不耐烦	impatient	紧张	stress
活跃	active	冷漠	indifferent	沉闷	tedious
自信	confident	烦躁	irritable	困倦	tired
兴奋	excited	模糊	vague	Affine	
流畅	fluent	着急	worried	亲和	affine
激情	passionate	Bored		和蔼	amiable
积极	positive	无趣	bored	轻松	easy
洪亮	resonant	冷静	clam	温婉	gentle
严肃	serious	迷茫	confused	开心	happy
语气生硬	stiff	呆板	dull	幽默	humorous
严厉	strictly	随意	free	活泼	lively
豪放	unrestrained	犹豫	hesitating	温和	mild
Impatient		平淡	insipid	耐心	patient
失望	disappoint	懒散	lazy	真诚	sincere
反感	disgust	安静	quite	温柔	soft

Figure 2: 41 Teaching Style Adjectives.

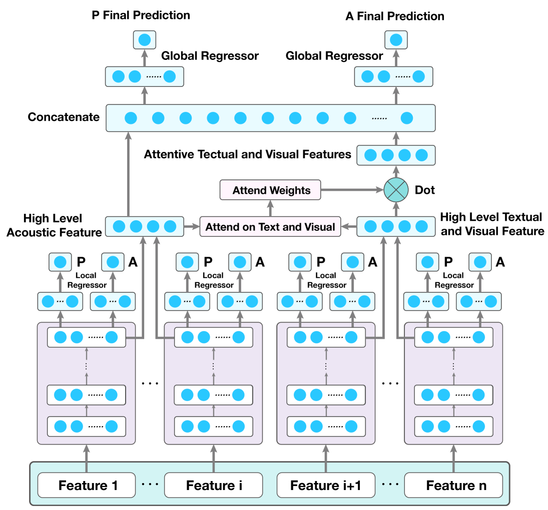

Figure 3: The structure of AMMDNN

right while negative adjectives like impatient, rigid and stiff lay in the bottom left.

4.2 Attention-based Multi-path Multi-task Deep Neural Network

Intuition. For traditional cross-media dimensional modelling, the entire features are employed as input and trained in a single regressor, which causes a high feature dimension. Thus, for a limited labeled training data, it would restrict the prediction performance to a great extent. Meanwhile, multi-task learning can improve generalization of a model by learning from related tasks [4][27]. Therefore, two strategies are applied to resolve our issue: 1) We adopt an attention-based Multi-path Deep Neural Network (MDNN) [28] to capture the internal correlations between cross-media features.First, it trains raw features from groups in local regressors to avoid high dimensions. Then high-level features of each local regressors of different modalities are concatenated based on attention mechanism as the input of a global regressor. More importantly, both local and global regressors are trained at the same time through a single objective function. 2) Intuitively pleasure and arousal which we predicted have close relations. Therefore, we promote the AMDNN

Notation	Definition
a	the high-level acoustic feature
w	the concatenated high-level visual and textual feature
w_i	the i-th dimensional feature of $w$ïijŇ
m	the dimension number of w
$f_a(\cdot, a)$	attention function
v	the attentive visual and textual feature
v_i	the i-th dimensional feature of v
x	the concatenated vector of a and v
y	the ground-truth label
$\hat{y}$	the prediction result
T	the number of tasks
N	the number of *local regressor*
$\mathcal{L}_{t,g}$	the cost function for *global regressor* and the t-th task
$\mathcal{L}_{t,l,n}$	the cost function for n-th *local regressor* and t-th task
λ	the weight coefficient that between 0 and 1

Table 1: Notations and Definitions.

to a novel structure named Attention-based Multi-task Multi-path Deep Neural Network (AMMDNN), shown in Figure 3.

The structure of AMMDNN. Instead of learning a single regressor with the whole sample features, the raw features are divided into small groups to learn multiple regressors based on different low-level descriptors (LLDs) and statistical functions, such as mean or standard deviation of MFCC features [28]. Therefore, each feature of different modalities can be used to train the responding regressor, which is called *local regressor*. With the approach, the problem of high-dimensional inputs can be effectively avoided.

It is noteworthy that although *local regressors* take the independent nature of features into account, they largely ignore the relationships between different groups and modalities. To solve this issue, we merge the highest hidden layers of each *local regressors* to generate a global representation based on the attention mechanism. Specifically, to better take advantages of the visual information and textual information, we adopt the attention mechanism mentioned in [5] and modify the highest hidden layers to concatenate high-level representation features.

For clarity, Table 1 illustrates several important notations and their definitions used in the paper. w_i, $i \in (1, m)$ represents the i-th dimensional feature of concatenated high-level visual and textual feature w, and $f_a(\cdot, a)$ denote the attention function conditioned on the current high-level acoustic feature$a$. v_i represent the i-th dimensional feature of the attentive visual and textual feature v. Then, the attention weight α_i and attentive visual and textual feature v_i is formulated as follows:

$$u_i = f_a(w_i, a) \tag{2}$$

$$\alpha_i = \frac{exp(u_i)}{\sum_{i=1}^{m} exp(u_i)} \tag{3}$$

$$v_i = \alpha_i \cdot w_i \tag{4}$$

A fully-connected layer with Exponential Linear Units (ELU) [6] activation is chosen as the attention function and the attention vector v is concatenated with the high-level acoustic feature a as the new input of the global regressor. Thus the concatenated high-level

Regression	P RMSE	P CCC	A RMSE	A CCC
DNN	0.844	0.554	0.722	0.669
SVM	**0.680**	0.703	**0.565**	0.787
DTree	1.006	0.511	0.834	0.623
RF	0.715	0.672	0.596	0.760
MDNN	0.707	0.694	0.648	0.765
AMDNN	0.710	0.729	0.587	0.805
AMMDNN	0.713	**0.739**	0.582	**0.808**

Table 2: Comparison among different regression models for arousal and pleasure prediction.

representation vector x becomes $[a, v]$. Then, this representation is applied to train a *global regressor*, which can effectively detect the correlations between different groups of features. Moreover, we optimize the framework through the single objective function, as shown in Equation 7. The *local regressors* and *global regressor* are trained simultaneously with the function.

Meanwhile, [4] and [27] show that multi-task learning can improve model generalization by learning from related tasks. Intuitively predicting pleasure and arousal values have close relations[4]. Therefore, we can utilize multi-task learning to predict pleasure and arousal values together. Specifically, each regressor (both local and global ones) has two prediction tasks.

The mean square error (MSE) is applied as the loss function, which minimizes:

$$\mathcal{L} = (\hat{y} - y)^2 \tag{5}$$

where y is the ground-truth label, $\hat{y}$ is the prediction result. Therefore, our total loss function for the Multi-path Multi-task Deep Neural Network is as follows:

$$\mathcal{L} = \sum_{t=1}^{T}((1-\lambda)\mathcal{L}_{t,g} + \lambda \sum_{n=1}^{N} \mathcal{L}_{t,l,n}) \tag{6}$$

Across all tasks T, $\mathcal{L}_{t,g}$ is the cost function for *global regressor* and the t-th task while $\mathcal{L}_{t,l,n}$ is the cost function for n-th *local regressor* and t-th task. λ is the weight coefficient that between 0 and 1(we set λ=0.5 in our experiment). N is the number of *local regressor* and T is the number of our tasks.

As shown in Figure 3, the motivation of this acoustic-guide Attention-based Multi-path Multi-task Deep Neural Network is that we use the acoustic features to guide the attention weights of the textual features and visual features in order to enforce the model to self-select which dimension feature it should attend on.

4.3 Mapping Coordinates with Teaching Styles.

Based on the proposed AMMDNN, we can get two-dimensional coordinates for each utterance. Then we select the adjective word on the TSSS with the shortest Euclidean distances for the coordinates. In this way, we can obtain three teaching style words that fit with the utterance most.

5 EXPERIMENTS

5.1 Data Collection

TEG18. We build a sizeable full-annotated benchmark dataset, which entirely contains 4,541 cross-media utterances recorded in the educational environment from TAL Education Group(TEG18). All of the cross-media utterances are recorded from the real primary school classes. Data samples are randomly collected from different teachers and different courses. The dataset covers both male and female teachers. Specifically, there are 42 teachers in total, and 24 are female while 18 are male. Also, each of the cross-media utterances is 10 seconds long. An example of data in TEG18 is shown in Figure 5. , and the main contents are teachers' speeches.

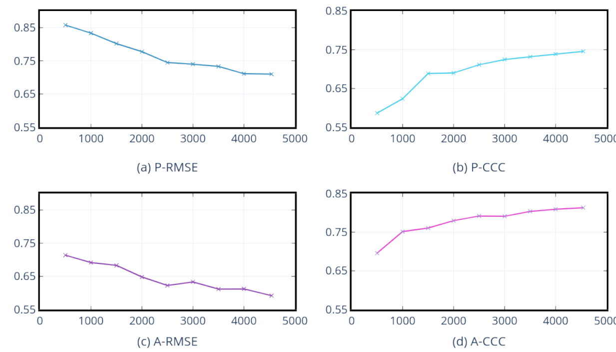

Figure 4: Training data scalability analysis.

	-2	-1	0	1	2	
Wide awake						Sleepy
Friendly						Strict
Harsh Sound						Comfortable Sound
Cautious						Surprising

Table 3: Four questions for the labelers to answer.

5.2 Labeling

In terms of pleasure and arousal labeling, based on the PAD model proposed by [17], [14] provides a Chinese version questionnaire to evaluate pleasure, arousal, and dominance values. Specifically, annotators need to answer four different questions for each value. Then, the paper gives three formulas for calculating the three values.

For our work, considering the data are recorded in the teaching environment, the dominance of teachers will be high, we annotate the teaching styles adjectives with pleasure and arousal labels. Based on the Chinese version questionnaire, we make some changes and reduce the number of questions to eliminate the ambiguity and make the survey more suitable for the education environment. As is shown in Table 3, we design four questions for annotators to answer. Specifically, question 1 and 4 are about arousal value and question 2 and 3 are about pleasure value. For each question, there are two different descriptions indicating -2 and 2 respectively. Annotators need to choose a number between -2 and 2. Then, we normalize all the labeled values to make the mean 0 and the variance 1.

$$P = -Q_2 + Q_3 \tag{7}$$

$$A = -Q_1 + Q_4 \tag{8}$$

Each label of every utterance is annotated by four different annotators independently. To access inter-rater reliability, we calculate the Cronbach alpha coefficients for all labelers regarding pleasure and arousal values. The Cronbach alpha coefficients are 0.782 for pleasure and 0.828 in terms of arousal values. Compared with the public PAD full-annotated dataset IEMOCAP [3], where Cronbach alpha coefficients are 0.809 and 0.607 in terms of pleasure and arousal values, the results prove the effectiveness of our dataset.

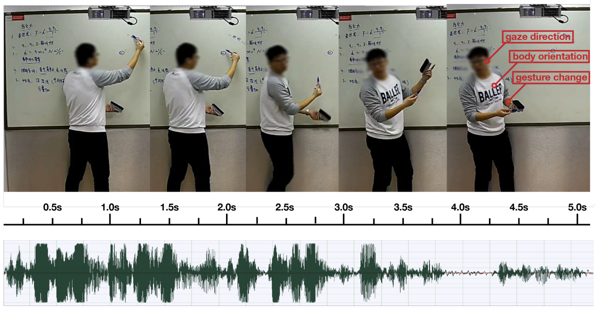

Figure 5: An example of data in TEG18.

Visual Features	Degree	Description
Teachers' facial expression	3	especially obvious laugh;
	2	obvious smile;
	1	serious;
	0	else;
Body and eyes orientation	3	the teacher is in the face of student most of the time, but occasional turning head or turning back (within two times);
	2	the teacher 's sight moves between the blackboard and the student (more than two times);
	1	most of the time, the teacher turns his/her back to the students, has no eye contact or eye contact is not frequent;
	0	else;
Posture change	2	excited: the hand movement amplitude is large, or the teacher's hand has been moving or move fast;
	1	not excited: the hand is almost motionless or writing in a small area, or can't be seen the whole time in the video;
	0	other: always draw graphics on the blackboard, etc.
Interaction	2	One or more students raise their hands, stand up or make other gestures;
	1	The teacher keeps talking or tries to communicate with the students;
	0	else;

Figure 6: Visual feature details.

5.3 Feature Extraction

Acoustic feature. We utilize openSMILE toolkit [8] to extract acoustic features for TEG18. Totally, we obtain 1,582 statistic acoustic features, which is the same as the acoustic features used in the INTERSPEECH 2010 Paralinguistic Challenge [19].

Visual feature. We define the visual features of teachers' teaching classes from four aspects: degree of teachers' facial expression, degree of body and eyes orientation, degree of posture change, degree of interaction. We invite three people to annotate them, and the annotation details are listed in Figure 6. Totally, we have 14-dimensional visual features.

Textual feature. For textual information in TEG18, we first use Thulac Tool[15] which is an efficient Chinese word segmentation to get words of an utterance. Then we utilize word2vec to learn word embeddings. Specifically, we use the whole 31.2 million chinese word corpora collected from the 7.5 million utterance from SVAD13[28] as the training corpora for word2vec. Then, we extract 4200-dimensional utterance-level textual features according to the statistic functions (mean, std, disp, max, min, range, quartile1/2/3, iqr1-2/2-3/1-3, skewness, kurtosis) over the LLDs.

5.4 Experimental Setup

Evaluation metrics. In all the experiments, we evaluate the performance in terms of concordance correlation coefficient(CCC). The results reported in TEG18 are based on five-fold cross validation.
Comparison methods. For predicting arousal and pleasure values, we utilize six different baseline regression models: Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DTree) [25], Deep Neural Network (DNN) [2], Multi-path Deep Neural Network (MDNN)[28], and Attention-based Multi-path Deep Neural Network(AMDNN).

Construction setting. In implementation of comparisons, we set λ=0.5, N = 7 in Eq. 7 for the comparison methods utilize multi-path structure. Each local regressor contains two hidden layers with 400 units. Dropout [22] is applied for each hidden layer with a dropout ratio of 0.5. The optimization method we adopt is Adam [12] with an initial learning rate at 10^{-4}. And the activation function we apply is ELU activation.

5.5 Performance

To evaluate the effectiveness of our proposed Multi-path Multi-task Deep Neural Network(AMMDNN), we compare the performance with some baseline methods with cross-media information. Table 2 shows the results. In terms of CCC, the proposed AMMDNN outperforms all the baseline methods: for arousal, +0.185 compared with DTree, +0.048 compared with RF and +0.021 compared with SVM. For pleasure, +0.228 compared with DTree, +0.067 compared with RF and +0.036 compared with SVM. Specifically, 1) the MDNN outperforms the DNN (+0.096) in terms of CCC with arousal and (+0.140) in terms of CCC with pleasure. This proves the effectiveness of the multi-path component which considers the in-dependency nature of different features and avoid high-dimensional inputs with limit training data . 2) the AMDNN outperforms the MDNN (+0.040) in terms of CCC with arousal and (+0.035) in terms of CCC with pleasure. This proves the effectiveness of the proposed attention mechanism, which utilize the acoustic features to guide the attention weights of the textual features and visual features to help enforce the model to self-select which dimension feature it should attend on. 3) the AMMDNN outperforms the AMDNN (+0.003) in terms of CCC with arousal and (+0.010) in terms of CCC with pleasure. This proves the effectiveness of the proposed multi-task method which consider predicting pleasure and arousal values as related tasks.

5.6 Analysis

Feature contribution analysis. We first discuss the contributions of acoustic, visual and textual features in understanding teaching styles. Specifically, for 'Acoustic Only', 'Acoustic+Visual', 'Acoustic+Visual+ Textual', we utilize Multi-path multi-task Deep Neural Network model. And for 'Acoustic+Visual+Textual', we utilize AMMDNN model. The CCC results for arousal and pleasure are shown in Figure 7. As show in the figure, the "Acoustic+Visual" outperforms the "Acoustic Only" (+0.014) in terms of CCC with

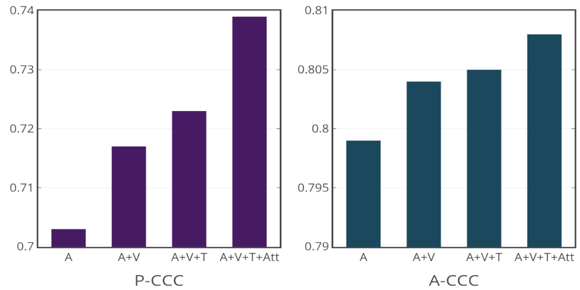

Figure 7: Feature contribution analysis.
(A:Acoustic, V:Visual, T:Textual, Att:Attention)

pleasure and (+0.005) in terms of CCC with arousal. This validates the necessity of taking the textual information into consideration. The "Acoustic+Visual+ Textual" outperforms the "Acoustic+Visual" (+0.006) in terms of CCC with pleasure and (+0.001) in terms of CCC with arousal, which validates the necessity of taking the textual information into consideration. The "Acoustic+Visual+Textual+Attention" outperforms the "Acoustic+Visual+ Textual" (+0.016) in terms of CCC with pleasure and (+0.003) in terms of CCC with arousal, which indicates that our proposed attention mechanism can be more effective in modelling multi-modal features.

Parameter sensitivity analysis. We further test the parameter sensitivity about training data size. From Figure 4, we can find that as the scale of training data increases, performance obviously gets better for all of the evaluation metrics (pleasure RMSE, pleasure CCC, arousal RMSE, and arousal CCC). Specifically, the pleasure has a higher improvement than arousal. The growth trend is slowing down after training size reaches 2500. With the size over 4500, the performance almost reaches convergence. Considering the difficultly for manually labelling, we choose 4541 as our dataset.

5.7 Case Study

With the proposed AMMDNN, we are capable of mapping cross-media features to two-dimensional coordinates on the TSSS. Then, we select the adjective word on the TSSS with the shortest Euclidean distances for the coordinates so that we can get the most suitable teaching style word for each utterance. Based on our prediction results for teaching style adjectives, we conduct some interesting case studies to further show the advantages and universality of our solution.

Teaching styles comparison among different teachers. We can establish a personal TSSS for every teacher to analyze his/her teaching styles. To determine the teaching styles of each teacher, first, we apply AMMDNN to calculate the pleasure and arousal values of an utterance set corresponding to the teacher. Then, we assign an coordinate value for the teacher by calculating the center of gravity of the points in the utterance set. Finally, we choose three teaching style words which have the shortest Euclidean distance with our calculated teacher coordinate value in TSSS. Therefore, the three chosen teaching style words are considered as the teaching style for the teacher.

We randomly choose two male teachers shown in Figure 10.(a) and (b) and two female teachers shown in Figure 10.(c) and (d). Based on the TSSS, the teacher in Figure 10.(a) may have the teaching styles of insipid, dull and quite, since he has high pleasure values and quite low arousal values. In terms of the teacher in Figure 10.(b), this teacher has low pleasure and high arousal values. Therefore, he may have the teaching styles of serious, resonant and unrestrained. Regarding the teacher in Figure 10.(c), she has a wide range of the pleasure and arousal values, while the values are all not very high, indicating that she may have sincere, humorous and lively teaching styles. The teacher in Figure 10.(d) may be sincere, humorous and lively, since she has high pleasure and arousal values. Based on the TSSS, we can better understand teachers' teaching styles and analyze them objectively and give them quantitative feedbacks.

Teaching styles comparison among different courses. We further analyze teaching styles among different courses. We apply the proposed AMMDNN to predict both pleasure and arousal values for different courses. By drawing the pleasure-arousal values on our TSSS for each course, a variety of teaching style features among different courses is revealed evidently. Take physics as an example, it (Figure 8, graph(a)) appears a considerable tendency toward lower pleasure value when compared with Chinese(Figure 9, graph(a)), which indicates that in a class for science subjects such as physics, the teacher generally acts more seriously than in a class for subjects that do not put forward such a strong request for scientific thinking.

Furthermore, we also analysis the change of teaching style over different periods in a class for certain subject. We divide each class into five segments by time equally. Then we draw the data in different segment separately in our TSSS, as shown in Figure 8.(b-f) and 9.(b-f). For each segment, we also compute the average pleasure and arousal value and show it on our TSSS. And the outcome graphs display distinct teaching style changes by time in different courses. Again, the physics course is treated as an example. As shown in 8.(b-f), with the timeline of a physics class proceeds, the pleasure value start to arise and arousal value start to shrink. And that phenomenon sits well with the common sense that in such a scientific class which is usually accused of excessive seriousness, the teacher gradually becomes more exhausted and less enthusiastic than at the beginning of the class.

Teaching styles analysis among different attention rate during a class. We calculate students' concentration on teachers during a class and analyze the results with the predicted pleasure and arousal values.

First, we calculate how much the students pay attention to the teacher. For each frame of the teaching class video, we calculate the students' head position coordinates and the students' line of sight vector in the same coordinate system. First, the head coordinates and the line of sight vector are projected onto a two-dimensional xy plane. Second, we calculate all intersections of the students' line of sight and record the number of intersections as n. Third, we sort the x coordinate values of the intersections and calculate the average after removing the largest $n/4$ values and the smallest $n/4$ values. This average is regarded as the predicted x coordinate of the teacher. Similarly, we can get the predicted teacher's y coordinate. Finally, we get the distance from the predicted position of the teacher on the xy plane to the line of sight for every student. Then, we divide them into three categories(category I, II, III) according to the distance mention above. Specifically, scores under half of the average are put into category I, which means a lack of attention. Scores more

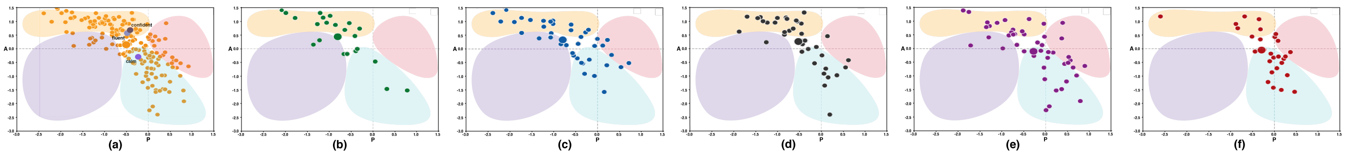

Figure 8: Physics course styles analysis.

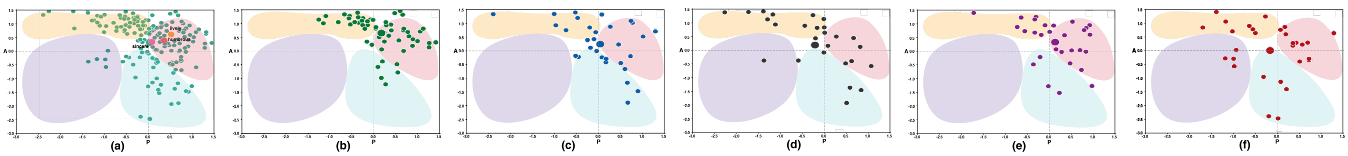

Figure 9: Chinese course styles analysis.

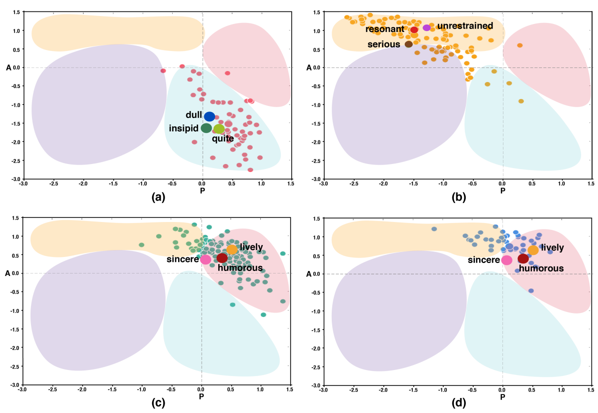

Figure 10: Personal teaching styles analysis.

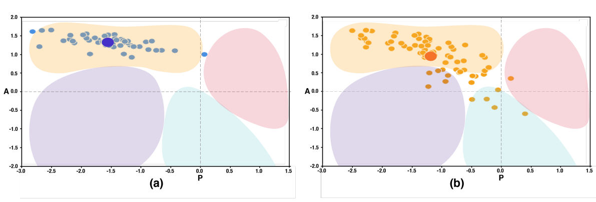

Figure 11: Teaching styles analysis among different attention rate for lesson(I).

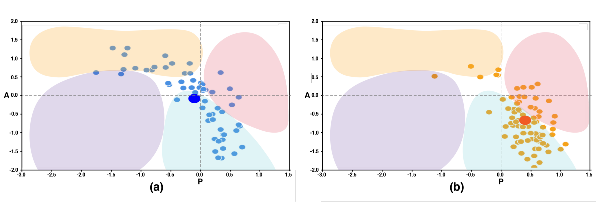

Figure 12: Teaching styles analysis among different attention rate for lesson(II).

than twice the average are put into category II, which means the teacher is highly focused on. Otherwise, it will be put into category III.

We collect a dataset containing 39 lessons from real high school teaching classes provided by EduBrain.ai. Then we randomly choose two lessons and cut them into 10s utterances. Next we predict the pleasure and arousal values of these utterances utilizing our proposed model. As shown in Figure 11 and 12, Figure 11.(a) and 12.(a) are utterances of category II, which have high student attention. Figure 11.(b) and 12.(b) are utterances of category I, which show low attention of students. We suggest that students pay more attention when teachers have lower pleasure values and higher arousal values in class. Therefore, teaching styles like passionate and severe which have low pleasure and high arousal values may catch more attention in class.

6 CONCLUSION

In this paper, we make an important step towards understanding teaching styles. We build a Teaching Styles Semantic Space (TSSS) to describe teaching styles. Then, we select the most often used 41 teaching style adjectives from more than 10,000 feedback questionnaires provided by Tomorrow Advancing Life (TAL) Education Group. Further, we propose an Attention-based Multi-path Multi-task Deep Neural Network (AMMDNN) to map cross-media features to the coordinates on the TSSS. Our result shows that AMMDNN turns out to be a practical solution. Based on the TSSS, we can obtain the teaching style adjective word that have the shortest Euclidean distances with the coordinates so that we can finally map teachers' utterances with teaching styles. Moreover, we conduct some interesting case studies which show the effectiveness of the dimensional space for understanding teaching styles.

ACKNOWLEDGMENTS

This work is guided by Mr. Yan Huang from TAL Education Group. We would like to sincerely thank TAL Education Group and Mr. Yan Huang for their support and guidance to the work. We would also like to thank EduBrain.ai and Shijiazhuang No.1 High School Hebei for their support.

This work is supported by National Key Research and Development Plan (2016YFB1001200),the Innovation Method Fund of China (2016IM010200), the state key program of the National Natural Science Foundation of China (NSFC) (No.61831022) and National Natural, and Science Foundation of China(61521002).

REFERENCES

[1] Madeleine Atkins and George Brown. 2002. *Effective teaching in higher education.* Routledge.

[2] Yoshua Bengio. 2009. Learning Deep Architectures for AI. *Foundations & TrendsÂő in Machine Learning* 2, 1 (2009), 1–127.

[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.

[4] Shizhe Chen, Qin Jin, Jinming Zhao, and Shuai Wang. 2017. Multimodal multitask learning for dimensional and continuous emotion recognition. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge.* ACM, 19–26.

[5] Yuxiao Chen, Jianbo Yuan, Quanzeng You, and Jiebo Luo. 2018. Twitter Sentiment Analysis via Bi-sense Emoji Embedding and Attention-based LSTM. In *2018 ACM on Multimedia Conference.* ACM, 117–125.

[6] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2015. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289* (2015).

[7] Florian Eyben, Bjorn Schuller, Stephan Reiter, and Gerhard Rigoll. 2007. Wearable assistance for the ballroom-dance hobbyist-holistic rhythm analysis and dance-style classification. In *2007 IEEE International Conference on Multimedia and Expo.* IEEE, 92–95.

[8] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia.* ACM, 1459–1462.

[9] Joan Gorham. 1988. The relationship between verbal teacher immediacy behaviors and student learning. *Communication education* 37, 1 (1988), 40–53.

[10] Joe E Heimlich and Emmalou Norland. 1994. *Developing Teaching Style in Adult Education. The Jossey-Bass Higher and Adult Education Series.* ERIC.

[11] Joe E Heimlich and Emmalou Norland. 2002. Teaching style: where are we now? *New directions for adult and continuing education* 2002, 93 (2002), 17–26.

[12] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[13] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee. 2003. Emotion recognition by speech signals. In *Eighth European Conference on Speech Communication and Technology.*

[14] Xiaoming Li, Haotian Zhou, Shengzun Song, Tian Ran, and Xiaolan Fu. 2005. The reliability and validity of the Chinese version of abbreviated PAD emotion scales. In *International Conference on Affective Computing and Intelligent Interaction.* Springer, 513–518.

[15] Zhongguo Li and Maosong Sun. 2009. Punctuation as implicit annotations for Chinese word segmentation. *Computational Linguistics* 35, 4 (2009), 505–512.

[16] Albert Mehrabian. 1980. Basic dimensions for a general psychological theory implications for personality, social, environmental, and developmental studies. (1980).

[17] Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology* 14, 4 (1996), 261–292.

[18] Gelareh Mohammadi and Alessandro Vinciarelli. 2011. Humans as feature extractors: combining prosody and personality perception for improved speaking style recognition. In *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on.* IEEE, 363–366.

[19] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan. 2010. The INTERSPEECH 2010 paralinguistic challenge. In *Proc. INTERSPEECH 2010, Makuhari, Japan.* 2794–2797.

[20] Jialie Shen, John Shepherd, Bin Cui, and Kian-Lee Tan. 2009. A novel framework for efficient automated singer identification in large music databases. *ACM Trans. Inf. Syst.* 27, 3 (2009), 18:1–18:31.

[21] Rajka Smiljanić and Ann R Bradlow. 2009. Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and linguistics compass* 3, 1 (2009), 236–264.

[22] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

[23] Maosong Sun, Xinxiong Chen, Kaixu Zhang, Zhipeng Guo, and Zhiyuan Liu. 2016. *Thulac: An efficient lexical analyzer for chinese.* Technical Report. Technical Report.

[24] Nattapong Thammasan, Ken-ichi Fukui, and Masayuki Numao. 2017. Multimodal Fusion of EEG and Musical Features in Music-Emotion Recognition.. In *AAAI.* 4991–4992.

[25] Adam Trendowicz and Ross Jeffery. 2014. *Classification and Regression Trees.* Springer International Publishing. 1174âĂŞ1176 pages.

[26] Alan H Wurtzel and Joseph R Dominick. 1971. Evaluation of television drama: Interaction of acting styles and shot selection. *Journal of Broadcasting & Electronic Media* 16, 1 (1971), 103–110.

[27] Biqiao Zhang, Emily Mower Provost, and Georg Essi. 2016. Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on.* IEEE, 5805–5809.

[28] Suping Zhou, Jia Jia, Qi Wang, Yufei Dong, Yufeng Yin, and Kehua Lei. 2018. Inferring Emotion from Conversational Voice Data: A Semi-supervised Multi-path Generative Neural Network Approach. (2018).