A Study of Classifying Style of Teachers & State of Students’ Learning based on K12 Online Education

Tuanji GONG\textsuperscript{a,*}, Xuefeng ZHENG

School of Computer and Communication Engineering, University of Science and Technology Beijing, 100083, China

\textsuperscript{a}email:gongtuaji@foxmail.com

Keywords: Speaker Diarization, Speaker Recognition, Style of Teachers, State of students’ Learning, Deep Learning, Classification, K12

Abstract. In recent years, online education has been advancing significantly. However there is a major challenge how to evaluate style of teachers and state of student learning. In this paper, we propose a novel method that combines speaker diarization, speaker recognition, feature selection to classify style of teachers and state of students learning based on audio data. We train speaker recognition model and learn embedding vector of teachers or students on online platform. We select 25 acoustic features and statistical features from audio recordings and train classification model to classify style of teachers and state of students’ learning jointly. Experimental results show that the task of classifying style of teachers achieves 71.25\% precision and precision of classifying state of students’ learning is 83.71\%.

Introduction

With the advent of Internet online education has broadly developed and emerged many famous organizations such as Khan academic, MOOC etc. Compared to traditional education, online education has the advantage of implementing personal education. Using internet and video technology, online education allows face to face communication virtually between teacher and student and improves efficiency. However, a huge challenge online education faces is how to evaluate style of teacher or to evaluate state of students learning. The principle goal of online education, especially K12 stage, is to improve students’ performance. Audio recording contains rich information and receives a lot attention. Recent many recent research focus on modeling teachers from audio recording or speech. Online education platform collects a lot of audio recordings and video recordings. However, as our knowledge, there is not any research on classifying style of teachers or state of students’ learning based on audio recordings.

With exponential growth of computer power and data volume and algorithm breakthrough, deep learning has achieved remarkable success in many fields [1]. In some fields, system performances such as speaker recognition[2], speech recognition[3] approach to human performance or amount to human level. Due to deep learning technology, deep speaker model achieves state of the art in speaker identification. Exploiting deep learning model achieves more accuracy of speaker identification and precision of segmenting between teacher and student.

In this work, we use convolutional neural network framework to train speaker identification model based on open source data set LibriSpeech [4] and ST-CMDS-20170001_1. Similar to framework in [5], but output dimension size in our model shrinkages to 256, reducing number of parameter. Using trained speaker identification model, we learn embedding vector of teacher or student according to their audio utterance in our platform and produced feature dataset of teacher or student. We process audio recordings and segment into audio segments and extract 25 interactive features, we take 25 features as input to train two classification model jointly: one is to classify style of teachers, another to classify state of the student learning.
Related Work

Recently, the research based on audio recording in live classroom has become a hot topic, attracting a lot of attention[6],[7],[8]. Blanchard et al.[9] evaluated a variety of automatic speech recognition(ASR) engines such as Google Speech and Bing Speech for automatic classroom dialog analysis. Some researcher design new multimodal method for data collection and automated analysis teacher-student interactions in live classrooms to identify instructional activities(e.g. lecturing, discussion) and to assess the quality of dialogic instruction[7]. Donnelly et al.[10] investigated automatic analysis of teacher strategies from audio recordings collected in live classroom, and train supervised model to identify occurrences of five key instructional segments including Question & Answer, Procedures and Directions, Supervised Seatwork, Small Group Work and Lecture.

In past decade, speaker diarization, which solve the problem of “who spoke what and when did they speak” in a conversation, has achieved brilliant process[11]. Speaker diarization consists of speaker segmentation and clustering components. Early stage employs I-vector feature to train model [12]. Inspired by deep learning applied in speech recognition successfully, a few researcher employ deep learning to model speaker diarization[18,19][13-17]. Speaker recognition is identifying an individual speaker from a set of potential speakers. Similar to speaker diarization, speaker recognition exploits low dimension feature based on factor analysis or I-vector to recognize speaker. Recently, benefited from powerful presentment of deep learning, a few of research employ deep neural network to train and learn embedding represent of speaker [2],[18] and to model end-to-end speaker recognition[19, 20].


System Architecture

The system is made up of speaker diarization, speaker recognition, feature selection and classification components. The structure of our system is shown in Fig.1.

![Diagram of system architecture](image)

**Speaker Diarization.** Speaker diarization consists of three mainly components including Voice Active Detection(VAD), speaker segmentation and speaker clustering. VAD is the primarily pre-process of speaker segment and speaker clustering. After extracting and segmenting audio recording into frames with 25ms length, frames are processed by VAD method introduced in [24] to
classify whether the frame is speech or not. We use speech segment with 0.1s window length proposed in [25] to split frames into homogeneous segments discriminated, and then feed these segments into deep recurrent convolutional neural network proposed in [17] to learn speaker embedding.

**Speaker Recognition.** Speaker recognition is essential in confirming speaker’s identity in online education application. In our environment there has usually two different speakers (e.g. a teacher and a student) in audio recording. In order to learn embedding represent of speakers, we use deep speaker model to train neural network to convolutional neural networks. Similar to deep speaker model[5], the difference of our model is the size of output dimension reducing to 256 and has 5 residential block layers[26]. Speaker recognition consists of one input layer, five residential block layer,1 average layer ,1 full connection layer with 2048x256 and one triplet layer. Every residual block consists of two convolutional layer with 3x3 filter and 1x1 stride. Every block has an identical structure and skip connection is the identity map.

After training speaker recognition, we learn speaker embedding for every teacher or student in online platform and build speaker dataset for identifying speaker.

**Feature Selection.** In this section, features are processed and selected from audio recording of online education for classifying style of teachers and state of students learning. Features consist of acoustic features and statistical features. Acoustic features of speakers such as volume, pitch, formant reveal speaker’s trait [23].Acoustic features detailed in Table 1 are extracted by pyAudioAnalysis tool[24]. After segmenting audio data by speaker diarization model and recognizing speaker by speaker recognition model, statistical features are selected and calculated, including speech session, speech distribute, communication between the teacher and the student and so on. In statistical features, a concept of session is proposed.

![Table 1](image)

<table>
<thead>
<tr>
<th>No.</th>
<th>Feature Description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>Mean of energy of teacher</td>
<td>The mean of energy(the sum of squares of signal value) of the teacher in audio recording</td>
</tr>
<tr>
<td>a2</td>
<td>Mean of pitch of teacher</td>
<td>The mean of pitch of the teacher in audio recording</td>
</tr>
<tr>
<td>a3</td>
<td>Mean of energy of student</td>
<td>The mean of energy of the teacher in audio recording</td>
</tr>
<tr>
<td>a4</td>
<td>Mean of pitch of student</td>
<td>The mean of pitch of the teacher in audio recording</td>
</tr>
<tr>
<td>a5</td>
<td>Mean of Zero crossing rate of teacher</td>
<td>The mean of the rate of sign-change of the signal of teacher speech frame</td>
</tr>
<tr>
<td>a6</td>
<td>Mean of Zero crossing rate of student</td>
<td>The mean of the rate of sign-change of the signal of student speech frame</td>
</tr>
<tr>
<td>a7</td>
<td>Mean of Spectral Roll off of teacher speech frame</td>
<td>The mean of frequency below which 90% of the magnitude distribute of the spectrum over teacher speech frames</td>
</tr>
<tr>
<td>a8</td>
<td>Mean of Spectral Roll off of student speech frame</td>
<td>The mean of frequency below which 90% of the magnitude distribute of the spectrum over student speech frames</td>
</tr>
<tr>
<td>a9</td>
<td>RMS energy of teacher speech</td>
<td>The root-mean-square energy of all teacher speech frames</td>
</tr>
<tr>
<td>a10</td>
<td>RMS energy of student speech</td>
<td>The root-mean-square energy of all student speech frames</td>
</tr>
</tbody>
</table>

The session contains several segments of continuous speech frames and total time length of non-speech frames between two speech segments is less than 5s. The number of session in online class can represent interaction between the teacher and the student.

All features including acoustic features and statistical features are normalized to have zero mean and unit variance by using Eq.1

\[ x = \frac{x - \mu}{\sigma} \quad (1) \]
Where $\mu$ stands for mean and $\sigma$ for standard variance.

**Classification Model.** Classification model takes selected features from audio data as input and classifies style of teachers and state of students’ learning jointly.

**Model of Classifying Style of Teachers.** Style of teachers has a variety of categories[21]. Based on interaction manner between the teacher and students in classroom, some researcher classify style of teachers into five categories, such as wise, emotion, nature, humor, and technology [22]. In our model, following[22] type of teachers

Style, objective function is negative log function described as Eq.2 and Eq.3

$$L(\theta) = \prod_{i=1}^{N} \prod_{j=1}^{K} \phi(x_j)^{(y=j)}$$  \hspace{1cm} (2)

$$l(\theta) = - \log L(\theta) = - \sum_{i=1}^{N} \sum_{j=1}^{K} l(y=j) \log \phi(x_j)$$  \hspace{1cm} (3)

Where $N$ is number of samples, $K$ is number of style type of teachers, $K=5$. $\phi(x_j) = \frac{e^{y_j}}{\sum_{i=1}^{K} e^{y_i}}$ is softmax function.

**Model of Classifying State of Student Learning.** Classifying state of student learning is a task of binary classification to predict student learning as active state or passive state. The objective function is cross entropy as showed in Eq4.

$$l(\theta) = \sum_{i=1}^{N} y_i \log \sigma(x_i) + (1 - y_i) \log (1 - \sigma(x_i))$$  \hspace{1cm} (4)

Where $N$ is number of sample, $y_i$ is target value, $\sigma(x_i)$ is predicted value.

The features selected or processed in section 3.3 are input to supervise machine learning method to train the classification model.

**Experiments**

**Dataset.** Speaker Dataset. We use LibriSpeech[4] dataset to train speaker recognition model. LibriSpeech dataset contains 2477 speakers including 1281 male and 1196 female, and every speaker has 10-15 utterances. In order to improve accuracy, we collect and set up new dataset of Chinese speaker from online education platform. The dataset contains 1652 Chinese speakers and every speaker has 13 utterances averagely. Among speakers, gender ratio of male and female approximates 1:1 and speaker’s age ranges from 12 to 60.

**Audio Dataset.** We collect audio recordings of one-to-one online teaching in online education platform from March, 2017 to October, 2017. The condition for candidate teachers is that the number of students that the teacher has teach is great than 5, and for candidate students is that the number of class the student has learned is more than 5. By above condition we pick up 800 teachers and 1000 students. Every audio recording of online class is evaluated and classified by three special workers. Style of teachers is assigned one of five categories, and state of student learning is labeled as one of active or passive learning. Total 6382 audio recordings are evaluated and labeled.

**Model Training.**

**Training Speaker Recognition Model.** Our model is based on Deep Speaker[5] with a slight difference that size of output dimension reduces to 256. Input feature is Filter Bank(Fbank) for Fbank includes more integrated acoustic information than Mel frequency cepstral coefficients(MFCC). Audio recordings are processed and converted PCM data by FFmpeg tool, and then Fbank feature with 64 dimension is converted and normalized to have zero mean and unit variance by pyAudioAnalysis. We employ TensorFlow [27] on Geforce 1080 GPU to train the model. During training stage, we use stochastic gradient descent (SGD) with Adam[28] optimal strategy. The model is trained for 20 epochs with a mini-batch size of 64 normalized by Batch Norm[29] and using Dropout[30] to avoid overfitting. The output of model is 256-dimensional embedding vector of speaker. Datasets are split into train dataset (80%), validation dataset (10%), and test dataset (10%).

**Training Classification Model.** Audio recordings are processed by FFmpeg and segmented into a lot small segments. Small segments are feed into speaker diarization model to recognize number of speaker and to tag every segment with corresponding speaker. Speaker audio segment is input to speaker recognition model to learn embedding of the speaker, and to identify the speaker(e.g. which
teacher or student) from speaker dataset by computing cosine distance between two speaker’s embedding. Acoustic features are extracted and tagged with corresponding speaker and statistical features are calculated. These features as input supervised classification method to predict result. Our dataset contains 6382 recordings and is split into train dataset(80%), validation dataset(10%) and test dataset(10%). In our experiment, we use logistic regression method and random forest classification method in scikit-learn tool to train and test the model.

Experimental Result. **Result of Speaker Recognition.** In experimental dataset containing LibriSpeech dataset and our dataset, speaker recognition model achieves 86.13% precision, close to performance of ResCNN model in [5].

**Result of Classifying Style of Teachers.** We compare precision of logistic regress method and random forest method in our dataset. In random forest method, criterion of a split is Gini and max-features attribute takes SQRT parameter. We conduct and compare precision in variety of number of tree. The precision of logistic regression method achieves 69.24%, and precision amounts to 71.25% by random forest method with 500 trees showed in first column in Table 3. Top 10 importance features for style of teachers is shown in Fig.2 (a) part. The most importance feature for classifying style of teacher is No. f4 in Table 2.

**Result of Classifying State of the Student Learning.** As well as classification of style of teachers, we conduct two classification method to predict state of the student learning. The precision of using logistic regression method is 81.08% vs 83.71% of using random forest method with 400 trees showed in second column in Table 3. Top 10 importance features for state of students’ learning is shown in Fig.2 (b) part. The most importance feature for state of students’ learning is No. f10 in Table 2.

<table>
<thead>
<tr>
<th>No.</th>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>f1</td>
<td>Length of speech</td>
<td>Length of all speech frames in audio recording(s)</td>
</tr>
<tr>
<td>f2</td>
<td>Length of teacher speech</td>
<td>Length of all speech frames that the teacher speaks in audio recording(s)</td>
</tr>
<tr>
<td>f3</td>
<td>Length of student speech</td>
<td>Length of all speech frame that the student speaks in audio recording(s)</td>
</tr>
<tr>
<td>f4</td>
<td>Proportion of teacher speech</td>
<td>Proportion of teacher speech in all speech in audio recording</td>
</tr>
<tr>
<td>f5</td>
<td>Proportion of student speech</td>
<td>Proportion of student speech in all speech in audio recording</td>
</tr>
<tr>
<td>f6</td>
<td>Ratio of teacher vs student</td>
<td>Ratio that length of speech teacher vs student</td>
</tr>
<tr>
<td>f7</td>
<td>Number of sessions</td>
<td>Number of session in audio recording</td>
</tr>
<tr>
<td>f8</td>
<td>Mean of session length</td>
<td>Mean of length of all sessions in audio recording</td>
</tr>
<tr>
<td>f9</td>
<td>Number of teacher session</td>
<td>Number of session launched by the teacher</td>
</tr>
<tr>
<td>f10</td>
<td>Number of student session</td>
<td>Number of session launched by the student</td>
</tr>
<tr>
<td>f11</td>
<td>Proportion of teacher session</td>
<td>Proportion of session launched by the teacher in all session</td>
</tr>
<tr>
<td>f12</td>
<td>Proportion of student session</td>
<td>Proportion of session launched by the student in all session</td>
</tr>
<tr>
<td>f13</td>
<td>Ratio of session</td>
<td>Ratio of session launched by teacher vs student</td>
</tr>
<tr>
<td>f14</td>
<td>Interval Mean of sessions</td>
<td>Mean of interval between two sessions</td>
</tr>
<tr>
<td>f15</td>
<td>Mean of interval of teacher and student speech</td>
<td>Mean of interval between teacher speech and student speech in audio recording</td>
</tr>
</tbody>
</table>
Table 3. Precision of two classification tasks

<table>
<thead>
<tr>
<th>Classification method</th>
<th>Precision of style of the teacher (%)</th>
<th>Precision of state of the student learning (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>69.24%</td>
<td>81.08%</td>
</tr>
<tr>
<td>Random forest</td>
<td>71.25%</td>
<td>83.71%</td>
</tr>
</tbody>
</table>

Conclusion

In this paper, we present a novel approach that classify style of teachers and state of students’ learning based on audio recordings. We train speaker recognition model based on Deep Speaker on LibriSpeech dataset and our dataset and build embedding database of students and teachers. The approach employs acoustic features and statistical features learning from audio recording by speaker diarization model and speaker recognition model and builds supervised classifier to style of teachers and state of the student learning. The experimental result shows that precision of the task classifying style of teachers is 71.25% and precision of classifying state of students’ learning achieves 83.71%.

Future work will focus on improving precision by introducing more features and more samples. We will also investigate an end-to-end approach to predict state of students’ learning signifdirectly.

References


