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

Tuanji GONG[a,*] , Xuefeng ZHENG

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

[a]email:gongtuanji@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].

Many researchers investigated style of teachers [21],[22].Yang et al.[22] researched correlation between style of teachers and style of students learning and classifies style of teachers into 5 categories: wise, emotion, nature, humor and technology. Hu et al.[23] researched the relationship between acoustic spectrum and personal trait.

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.

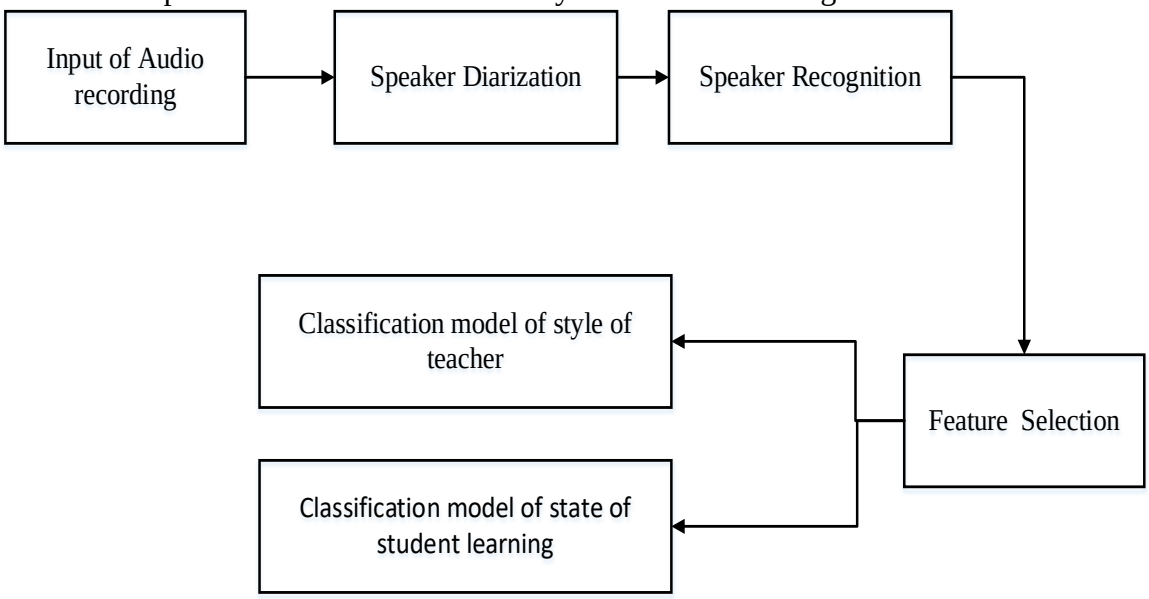

Fig.1. Diagram of system architecture

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 discriminately, 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. Acoustic features extracted from audio recordings of online class

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

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

$$x = \frac{x-u}{\sigma} \tag{1}$$

Where μ stands for mean and σ 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

$$\mathrm{L}(\theta) = \prod_{i=1}^{N} \prod_{j=1}^{K} \emptyset(x_j)^{I(y=j)} \tag{2}$$

$$\mathrm{l}(\theta) = -\log \mathrm{L}(\theta) = -\sum_{i=1}^{N} \sum_{j=1}^{K} I(y=j) log\emptyset(x_{j)} \tag{3}$$

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

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

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 a 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 2. Statistical features calculated from audio recordings of online class

No.	Feature	Description
f1	Length of speech	Length of all speech frames in audio recording(s)
f2	Length of teacher speech	Length of all speech frames that the teacher speaks in audio recording(s)
f3	Length of student speech	Length of all speech frame that the student speaks in audio recording(s)
f4	Proportion of teacher speech	Proportion of teacher speech in all speech in audio recording
f5	Proportion of student speech	Proportion of student speech in all speech in audio recording
f6	Ratio of teacher vs student	Ratio that length of speech teacher vs student
f7	Number of sessions	Number of session in audio recording
f8	Mean of session length	Mean of length of all sessions in audio recording
f9	Number of teacher session	Number of session launched by the teacher
f10	Number of student session	Number of session launched by the student
f11	Proportion of teacher session	Proportion of session launched by the teacher in all session
f12	Proportion of student session	Proportion of session launched by the student in all session
f13	Ratio of session	Ratio of session launched by teacher vs student
f14	Interval Mean of sessions	Mean of interval between two sessions
f15	Mean of interval of teacher and student speech	Mean of interval between teacher speech and student speech in audio recording

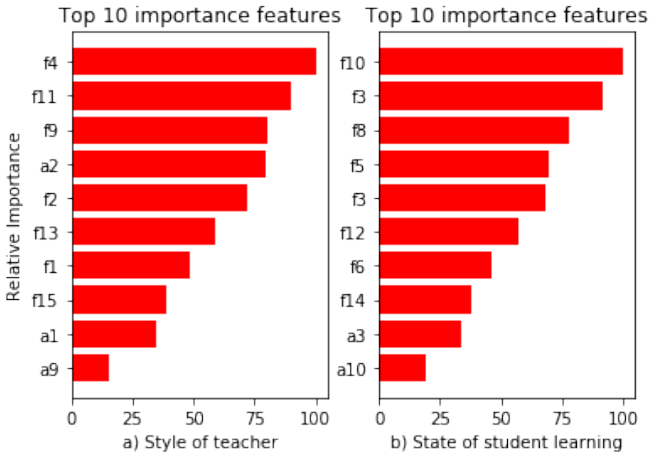

Fig.2. Import features for style of teachers & state of the student learning

Table 3. Precision of two classification tasks

Classification method	Precision of style of the teacher(%)	Precision of state of the student learning(%)
Logistic Regression	69.24%	81.08%
Random forest	71.25%	83.71%

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

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436-444, 2015.

[2] F. Richardson, D. A. Reynolds, and N. Dehak, "Deep Neural Network Approaches to Speaker and Language Recognition," IEEE Signal Processing Letters, vol. 22, pp. 1671-1675, 2015.

[3] L. W. W. Xiong , F. Alleva , Jasha Droppo , X. Huang , Andreas Stolcke, "The Microsoft 2017 Conversational Speech Recognition System," CoRR, vol. abs/1708.06073, 2017.

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in international conference on acoustics, speech, and signal processing, 2015, pp. 5206-5210.

[5] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, et al., "Deep Speaker: an End-to-End Neural Speaker Embedding System," 2017.

[6] N. Blanchard, S. D'Mello, A. M. Olney, and M. Nystrand, "Automatic Classification of Question & Answer Discourse Segments from Teacher's Speech in Classrooms," International Educational Data Mining Society, 2015.

[7] S. K. D'Mello, A. M. Olney, N. Blanchard, B. Samei, X. Sun, B. Ward, et al., "Multimodal Capture of Teacher-Student Interactions for Automated Dialogic Analysis in Live Classrooms," in ACM on International Conference on Multimodal Interaction, 2015, pp. 557-566.

[8] P. J. Donnelly, N. Blanchard, A. M. Olney, S. Kelly, M. Nystrand, and S. K. D'Mello, "Words matter: automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context," in International Learning Analytics & Knowledge Conference, 2017, pp. 218-227.

[9] N. Blanchard, M. Brady, A. M. Olney, M. Glaus, X. Sun, M. Nystrand, et al., A Study of Automatic Speech Recognition in Noisy Classroom Environments for Automated Dialog Analysis: Springer International Publishing, 2015.

[10] P. J. Donnelly, N. Blanchard, B. Samei, A. M. Olney, X. Sun, B. Ward, et al., "Automatic teacher modeling from live classroom audio," in Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, 2016, pp. 45-53.

[11] X. A. Miro, S. Bozonnet, N. W. D. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker Diarization: A Review of Recent Research," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, pp. 356-370, 2012.

[12] Y. Xu, I. V. Mcloughlin, Y. Song, and K. Wu, "Improved i-Vector Representation for Speaker Diarization," Circuits Systems and Signal Processing, vol. 35, pp. 3393-3404, 2016.

[13] H. Bredin and G. Gelly, "Improving Speaker Diarization of TV Series using Talking-Face Detection and Clustering," in acm multimedia, 2016, pp. 157-161.

[14] V. S. Ramaiah and R. R. Rao, "Speaker diarization system using HXLPS and deep neural network," Alexandria Engineering Journal, 2017.

[15] D. S. Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Alan McCree, "Speaker Diarization Using Deep Neural Network Embeddings," in ICASSP, New ORLEANS 2017, pp. 4930-4934.

[16] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. Mccree, "Speaker diarization using deep neural network embeddings," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2017, pp. 4930-4934.

[17] T. T. Paweł Cyrta, Wojciech Stokowiec, "Speaker Diarization using Deep Recurrent Convolutional Neural Networks for Speaker Embeddings," arxiv, 2017.

[18] V. Gupta, P. Kenny, P. Ouellet, and T. Stafylakis, "I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription," in international conference on acoustics, speech, and signal processing, 2014, pp. 6334-6338.

[19] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in Spoken Language Technology Workshop, 2017.

[20] P. G. D. Snyder, D. Povey, D. Garcia-Romero, Y. Carmiel, S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in IEEE Spoken Lang, 2016, pp. 165-170.

[21] Chun mei Zeng. A review of the teaching style of teacher. Continue Education Research. 2014(6): p. 141-144.

[22] Ligang Yang. A study of correlation on style of the teaching of teacher and style of the learning of student. Teaching & Administration. 2011. 2011(7).

[23] Chao Hu,Gengyue Fu. Recognize speaker by listening speech --- a research on audio spectrumand personality trait. Advances in Psychological Science. 2011. 19(6): p. 809-813.

[24] T. Giannakopoulos, "pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis," PLOS ONE, vol. 10, 2015.

[25] R. Wang, "Speaker Segmentation Using Deep Speaker Vectors for Fast Speaker Change Scenarios " presented at the ICASSP 2017.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in European Conference on Computer Vision, 2016, pp. 630-645.

[27] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," arXiv: Distributed, Parallel, and Cluster Computing, 2016.

[28] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," international conference on learning representations, 2015.

[29] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," international conference on machine learning, pp. 448-456, 2015.

[30] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, pp. 1929-1958, 2014.