Real-Time S-T Analysis Method Based on Speaker Recognition

Gui-lin LIU, Zeng-zhao CHEN* and Rong CHEN

National Engineering Research Center for E-Learning of Central China Normal University,
Hubei, Wuhan, 430079, China

*Corresponding author

Keywords: S-T analysis method, Speaker recognition, Class speech.

Abstract. With the continuous deepening of curriculum reform, teaching evaluation has become an important part of investigating the effectiveness of classroom teaching. This study improved the original S-T analysis method with speaker recognition technology that tracks the speaker's identity in class speech in real time. Combining the theoretical research of information technology-driven intelligent evaluation of classroom teaching, 6 lessons were selected as samples, and the original S-T analysis method and real-time S-T method were used for comparative study, which verified the effectiveness of the proposed real-time S-T analysis method.

Introduction

The new curriculum reform advocates the evaluation of "based on the process and promotes development", and puts forward new requirements on evaluation methods, tools, and evaluation implementation processes [1]. As the leader of teaching activities, the quality of classroom teaching is a direct reflection of teachers' teaching ability, and accurate evaluation of classroom teaching is an important information basis for improving teachers' teaching ability. Efficient teaching evaluation methods can provide practical guidance for teachers' teaching reflection and help teachers improve their teaching literacy. The traditional S-T analysis method uses on-site observation and manual recording, which is not suitable for large-scale classroom sampling and analysis. It cannot meet the needs of teachers for teaching reflection and professional development. Therefore, the improvement of the S-T analysis method is imminent.

The S-T analysis method is a classroom observation and research method, which can quantitatively analyze the classroom teaching process, estimate the teaching mode, and visualize the teaching process [2]. With the combination of technology and teaching, the classroom environment is gradually becoming more intelligent and modern, providing more possibilities and convenience for intelligent teaching evaluation [3]. This paper combined speaker recognition with original S-T analysis method, and proposed a real-time S-T analysis method, which is more accurate and efficient.

Related Works

ST analysis method is a quantitative analysis method to analyze teachers' teaching quality and teaching mode. It simplifies complex teaching behaviors into two types: teacher behaviors (T) and student behaviors (S). To improve the efficiency of classroom observation and the accuracy of data, behavior data is collected every 30 seconds. Finally, the classroom teaching model was analyzed according to the frequency of behavior conversion and teacher behavior occupancy, which provided a basis for teaching evaluation and theoretical research. However, the original S-T analysis method has four disadvantages: 1) the researcher needs to enter the classroom for observation, which is likely to cause interference to the teaching; 2) the on-site observation cannot be traced back, and the data is subject to subjective deviations; 3) Long interval sampling may cause data loss; 4) Manual observation is time-consuming and labor-intensive, which is not conducive to large-scale teaching analysis using this method.
With the development and application of information technology in the field of education, scholars have continued to improve the traditional S-T analysis. Some scholars improved the data collection method of the original S-T analysis method. For example, reference [4] improved S-T analysis method based on video. Due to the traceability of video analysis, the improved S-T analysis method used "time sampling and dynamic compensation", which effectively solved the problem of data loss. In a certain extent, the improved S-T analysis avoids the subjective bias of data collection, improves the accuracy of data collection and analysis, but it is still time-consuming and labor-intensive. Reference [5] combined image recognition with instructional video to recognize the teaching behavior of the sampling sequences through image recognition technology. Reference [6] distinguishes speaker identity by calculating the speech similarity of classroom audio clips to determine the S/T behavior of the sampled clips. These two methods make the analysis process more efficient and intelligent. However, the disadvantages of the low sampling accuracy of the original S-T analysis method still remain.

Method Description

The specific flow of the real-time ST analysis method proposed in this paper is shown in Fig.1. The microphone device is used to convert classroom teaching conversations into digital voice signals. Then the S/T behaviors are recognized in real time through speaker recognition. Finally, teacher behavior occupancy (Rt) and behavior conversion Rate (Ch) are calculated to support the teaching analysis. At the same time, the analysis results can also help teachers to reflect on teaching and improve teaching ability.

![Figure 1. Real-time S-T analysis method.](image)

Definition of S / T Behaviors

Different from the original S-T analysis method of fixed time interval sampling, the real-time S-T analysis method performs sampling based on behavior characteristics. According to the speaker recognition, the identity (teacher or student) of the speech are recognized in every second, so as to distinguish the time distribution of T behavior and S behavior. So the behavior definition of the improved S-T analysis method in this paper is shown in Table 1. Teachers' speech, explanation, questioning, and evaluation behaviors all take discourse as the carrier. The same is true for student responses, speeches, discussions, and so on. When students think, take notes, and complete classroom exercises, the classroom is silent. The discourse in the classroom is mainly composed of one teacher and several students. When using speaker recognition to judge S/T behavior, it is only necessary to determine whether the current utterance segment belongs to teacher. If it is, it is a T behavior; otherwise it is a student behavior.

<table>
<thead>
<tr>
<th>Category</th>
<th>Definition</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>T Behavior</td>
<td>Teacher's discourse behavior (such as lecture, question, evaluation, commentary)</td>
<td>Teacher’s speech</td>
</tr>
</tbody>
</table>
Speaker Recognition

Although the speaker's voiceprint feature distribution does not follow a specific distribution pattern, all distributions can be simulated by mixing multiple Gaussian models [7]. Therefore, the UBM-GMM model based on the Gaussian mixture model is the mainstream of speaker recognition. It uses a large number of non-target speaker’s speech data to train a background universal model (UBM) containing universal speakers' characteristics, then use a small amount of speech data of the target speaker to fit an exclusive probability distribution model on this UBM that conforms to its own voiceprint characteristics [8].

UBM is actually a high-order mixed Gaussian model, and its specific calculation formula is as follows:

$$P(X|\lambda) = \sum_{i=1}^{M} w_i b_i(X), \sum_{i=1}^{M} w_i = 1.$$  \hspace{1cm} (1)

Where $X$ is a high-dimensional speech feature vector, $b_i(X)$ is a single component of the Gaussian mixture model, $w_i$ is the weighting coefficient of the corresponding component, and $M$ is the number of components of the Gaussian mixture model. Each component in GMM is composed of the mean vector $\mu_i$ and the covariance matrix $\Sigma_i$, so the calculation method of each component is as follows:

$$b_i(X) = N(x, \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu_i)^T R_x^{-1} (x - \mu_i) \right\}. \hspace{1cm} (2)$$

In this study, a large number of different speakers’ speech data in the classroom environment are combined with the above formula and the EM algorithm to obtain the Gaussian mixture distribution of the UBM model. Classroom teaching is generally initiated by the teacher’s discourse. Therefore, this study uses the last minute speech data at the beginning of the class as standard teacher speech data, and uses the maximum posterior probability algorithm (MAP) to adaptively derive the teacher’s voiceprint model (GMM) on the UBM model.

Finally, for a given test voice data $O = (o_1, o_2, o_3, \ldots, o_t)$, a log-likelihood ratio is used to calculate the degree of matching between the voice data $O$ and the voiceprint model. If the similarity is high, it is the teacher's voice. Otherwise it is non-teacher voice. The specific calculation formula is as follows:

$$\text{Score}(O) = \log P(O|\lambda_{\text{GMM}}) - \log P(O|\lambda_{\text{UBM}}). \hspace{1cm} (3)$$

Data Process

The intelligent analysis method will improve the teacher behavior conversion rate (Ch) because the collected data is more complete and accurate. At the same time, due to the intersection between some teaching modes in the teaching mode structure distribution map of the original ST analysis, it is not conducive to the judgment of the results when the data falls at the intersection (see in Fig.2.(a)). Therefore, the teaching mode judgment standards need to be adjusted accordingly, and the threshold of the teacher behavior conversion rate (Ch) in the dialogue and mixed judgment standards is adjusted from 0.4 to 0.6 [9]. The specific judgment standards after adjustment are shown in Table 2 and Fig.2.(b). The calculation methods of Rt and Ch are shown in formula (4) and formula (5), respectively.

<table>
<thead>
<tr>
<th>S Behavior</th>
<th>Student utterances (such as answering, discussing, speaking)</th>
<th>Non-teacher’s speech</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Silence (such as thinking, taking notes, classroom exercises)</td>
<td>None speech</td>
</tr>
</tbody>
</table>
Figure 2. Structure of teaching model. (a) is the original structure, (b) is adjusted.

Table 2. Judgment criteria for teaching mode.

<table>
<thead>
<tr>
<th>Teaching model</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practicing</td>
<td>Rt ≤ 0.3</td>
</tr>
<tr>
<td>Lecturing</td>
<td>Rt ≥ 0.7</td>
</tr>
<tr>
<td>Dialogue</td>
<td>Ch ≥ 0.6</td>
</tr>
<tr>
<td>Mix</td>
<td>0.3 &lt; Rt &lt; 0.7, Ch &lt; 0.6</td>
</tr>
</tbody>
</table>

Rt = \frac{T \text{ Behaviors Duration}}{Total Class Duration}. \quad (4)

Ch = \frac{Behavior Conversion Times}{\sum (Single Behavior Duration/Average Behavior Duration)}. \quad (5)

**Experiment Analysis**

In this study, the effect of this method was tested in six primary school classrooms. A microphone was placed at the top of the center of the classroom, and the audio data was recorded as a mono wav audio file with a 16kHz sampling rate. The 13-dimensional Mel cepstrum coefficients of the audio data and its first-order difference and second-order difference are extracted, with a total of 39 dimensions. The experiment uses a 256-order Gaussian mixture model and sets the Score (O) threshold to 0. When the occurrence of utterances is detected, Score (O) is calculated. If Score (O) is higher than the threshold, the segment is judged as teacher utterance and marked as T behavior; otherwise it is student utterance and marked as S behavior. If no utterance appears, it is also marked as S behavior. The formula (6) is used to calculate the accuracy of the behavior recognition method, and the original S-T analysis method and the real-time S-T analysis method are used to estimate the teaching mode. The experimental results are shown in Table 3. It can be found that the real-time ST analysis system of this research has a high accuracy rate in identifying S/T behaviors. And it is basically consistent with the original ST analysis method in estimating the teaching mode. Therefore, the real-time ST analysis method can be applied to teaching analysis of daily classroom.

Accuracy of Behavior recognition = \frac{Correctly Recognized Behavior Duration}{Total Class Duration}. \quad (6)

Table 3. Experiment result.

<table>
<thead>
<tr>
<th>Class Number</th>
<th>Accuracy of S/T recognition</th>
<th>Rt</th>
<th>Ch</th>
<th>Teaching model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.934</td>
<td>0.5</td>
<td>0.64</td>
<td>Dialogue</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Original S-T</th>
<th>Real-time S-T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialogue</td>
<td>Dialogue</td>
<td>Dialogue</td>
</tr>
</tbody>
</table>
## Conclusion

In this study, discourse, which is the main carrier of information transmission in the classroom [10], was used as the main research object. The real-time S-T analysis method improved sampling method and behavior recognition method, which enables more accurate sampling and more efficient and intelligent real-time teaching process analysis. Real-time S-T analysis method can greatly improve the efficiency of teaching evaluation, is conducive to large-scale teaching analysis in daily classroom teaching. Hope it can have a positive meaning for teachers’ professional development and teaching theory research.

## Acknowledgement

Financially supported by the Fundamental research Funds for the Central Universities (grant no.CCNU19ZN002)

## References


