

# Adaptive Wavelength Calibration Algorithm for LAMOST

H. Q. Qin<sup>A</sup>, J. Zhu<sup>A</sup>, Z. Q. Zhu<sup>A</sup>, Z. F. Ye<sup>A,C</sup>, and A. L. Luo<sup>B</sup>

<sup>A</sup> Institute of Statistical Signal Processing, Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, P. R. China

<sup>B</sup> Department of LAMOST, National Astronomical Observatories of China, Beijing 100012, P. R. China

<sup>C</sup> Corresponding author. Email: yezf@ustc.edu.cn

Received 2009 July 16, accepted 2009 October 30

**Abstract:** An adaptive algorithm is presented for the wavelength calibration of the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST). The new algorithm can identify emission lines of calibration lamp without manual interaction and then fit relationship between the pixel positions and the wavelengths by utilizing a polynomial function. In this study, experiments are performed both on the actual data observed by LAMOST and the simulated data to analyze the accuracy and the robustness. The experimental results show that the proposed automatic algorithm can perform the lines identification exactly and acquire the wavelength solution accurately. Thus, an applicable approach is provided for the wavelength calibration of LAMOST.

**Keywords:** instrumentation: spectrographs — methods: data analysis — techniques: spectroscopic — line: identification

## 1 Introduction

The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) is a meridian reflecting Schmidt telescope with both large aperture and wide field of view (Cui 2009). The large focal plane of the telescope accommodates up to 4000 fibers, which are medially distributed to 16 multi-object fiber spectrographs. In each spectrograph, the spectra from fibers are split into blue (3700–5900 Å) and red (5700–9000 Å) channels by a dichroic filter, and finally focused onto 32 Charge Coupled Devices (CCD) with the size of  $4k \times 4k$  pixels. The large aperture, wide field of view and sufficient fibers promise a very high spectrum acquiring rate of more than ten-thousands of spectra per night, which presents a great challenge to the data reduction.

Wavelength calibration is an important step of the data reduction, which acquires the relationship between the CCD pixels and the wavelengths. The accuracy of wavelength calibration has a great influence on the subsequent data reductions, such as sky subtraction, redshift determination and so on.

Some wavelength calibration methods have been reported, such as the interpolation method (Bai et al. 2004), the linear regression method (Lo & Fountain 2006), and the polynomial fitting method (van Geffen & van Oss 2003; Sanchez 2006; Burles & Schlegel 2007; Thompson et al. 2009). However, these methods have some limitations in direct application to the wavelength calibration of LAMOST. This is because that they require either a manual interactive identification of the emission lines or

an initial wavelength solution with considerable accuracy. The details of the limitations are listed as follows:

- LAMOST can obtain 32 CCD frames with one exposure, and each frame has individual pixel–wavelength relationship. These relationships of different frames have discrepancies more or less.
- For a certain frame, the pixel–wavelength relationship will vary with the routine maintenance of the instruments and the unexpected external disturbances.
- The spectrum of the calibration lamp in red channel has considerable emission lines with significant discrepancy of intensity.
- The manual lines identification requires operator with rich experience.

As performing manual lines identification costs too much time and manpower, this method is unacceptable for LAMOST. For the alternate methods, different initial wavelength solutions should be preset to different frames and updated in a certain time, which is not convenient. Therefore, adaptive algorithms without manual interaction or initial wavelength solution have to be developed.

In this paper, an adaptive algorithm is developed for the wavelength calibration of LAMOST. The algorithm identifies the emission lines of the calibration lamp automatically, and then utilizes a polynomial function to fit the relationship between the CCD pixels and the wavelengths. In order to test the performance of the algorithm, experiments are implemented both on the observed spectra and simulated ones with artificial noise.

**Table 1. Emission lines of the calibration lamp in the red channel**

Wavelength <sup>a</sup>	Intensity <sup>b</sup>	Wavelength <sup>a</sup>	Intensity <sup>b</sup>
5769.6	700	6717.0	4200
5790.7	870	6929.5	5000
5852.5	740	7032.4	12000
5881.9	650	7174.0	850
5944.8	1900	7245.2	4800
5975.1	500	7439.0	1500
6030.0	550	7488.9	500
6074.3	2200	7535.8	400
6096.2	3900	8136.4	150
6143.1	7500	8300.3	300
6163.6	2000	8377.6	1400
6217.3	1500	8418.4	250
6266.5	4000	8495.4	800
6304.8	1400	8591.3	320
6334.4	5700	8634.7	300
6383.0	7000	8654.4	600
6402.3	20000	8680.8	200
6506.5	10000	8782.2	600
6532.9	2800	8853.9	150
6599.0	4000	8865.3	100
6678.3	6800		

<sup>a</sup>The unit of wavelength is Å.

<sup>b</sup>The intensities in the table are empirical relative values. For different spectra, they may vary more or less.

## 2 Wavelength Calibration Algorithm

### 2.1 The Spectra of the Calibration Lamp

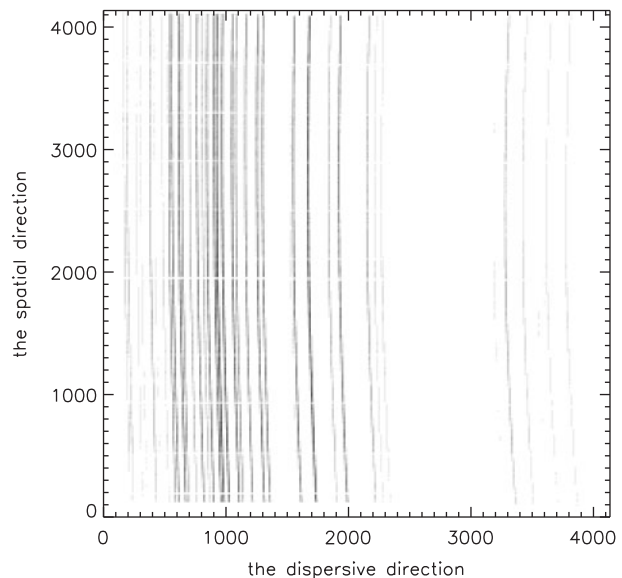
Before performing the wavelength calibration, an appropriate arc calibration lamp has to be selected. The selection of the lamp strongly depends on the number of the available emission lines and the coverage of the lines along the wavelength range. Considering the characteristics of LAMOST, a mercury vapor and neon lamp is selected for wavelength calibration. The wavelengths and intensities of the emission lines are listed in Table 1.

A two-dimensional (2D) spectral frame of the calibration lamp observed by LAMOST in the red channel is shown in Figure 1. The spectra from different fibers are distorted both in dispersive and spatial directions. One-dimension (1D) spectra are extracted from the frame by using the method reported in the authors' another article (Qin, Lee & Yuo 2009). A 1D spectrum extracted from a certain fiber is plotted in Figure 2, which shows that the intensities of different emission lines have great discrepancy.

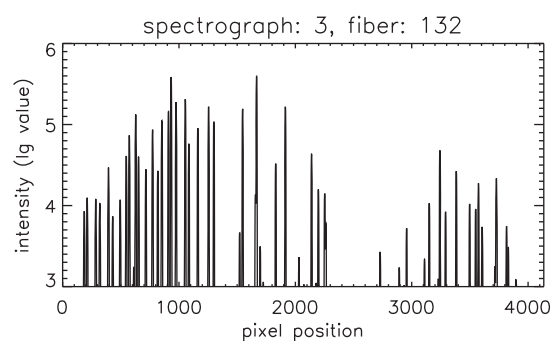
### 2.2 Identification of Emission Lines

The manual identification mainly depends on the wavelengths and the intensities of the emission lines, especially on the wavelength shifts and the relative intensities of the adjacent lines. The automatic algorithm is developed by imitating the manual interactive identification.

Because of the distortion shown in Figure 1, the pixel-wavelength relationship is nonlinear in the entire



**Figure 1** A 2D spectral frame of the calibration lamp observed by LAMOST in the red channel. The video format is reversed with black on white to improve the legibility.



**Figure 2** One certain 1D spectrum extracted from the spectra frame shown in Figure 1. The y-axis is set to log scale to enhance readability.

wavelength range. However, the pixel-wavelength relationship of the adjacent lines in a small region is approximately linear. The approximately linear relationship can be described as

$$\frac{\lambda_{i+1} - \lambda_i}{\lambda_i - \lambda_{i-1}} \approx \frac{x_{i+1} - x_i}{x_i - x_{i-1}}, \quad (1)$$

where  $\lambda_i$  is the wavelength of the  $i$ th line and  $x_i$  is the corresponding pixel coordinate illustrated in Figure 2. The relationship described in Equation (1) is an important reference of the manual identification.

The relative intensity of the adjacent emission lines is another available reference. The theoretical intensities are known from Table 1, and the actual intensities can be extracted from the observed spectra. As the efficiency of the instrument varies not too much in a small wavelength region, the theoretical values and corresponding actual values should have the following relationship:

$$\frac{V_t^{i+1}}{V_t^i} \approx \frac{V_o^{i+1}}{V_o^i}, \quad (2)$$

where  $V_t^i$  is the theoretical intensity of the  $i$ th line and  $V_o^i$  is the corresponding actual value.

The automatic identification algorithm is developed by imitating the manual identification. The wavelength of a certain line is invariable, but the intensity varies more or less from fiber to fiber. Therefore, the wavelengths are utilized as a primary identification standard while the intensities are considered as an assistant one. The automatic algorithm is performed in a four-step procedure as follows:

1. Locate possible emission lines.
2. Identify three initial lines.
3. Identify other lines iteratively.
4. Evaluate the identification and reject ill-identified lines.

### 2.2.1 Location of Emission Lines

The first step of the identification is to locate the possible emission lines of a certain spectrum; we refer the proposed method (Thompson et al. 2009) here. This method takes the derivative of the flux, and then looks for the positive to negative zero crossings with preset threshold. Afterwards, it performs a Gaussian fit to acquire exact centers and intensities of the emission lines. However, the method has a disadvantage in direct application because of the constant threshold. As shown in Figure 2, the intensities of different lines have great discrepancy. Some weak lines may be missed while the threshold is too large, and some noises may be mistaken as emission lines while the threshold is too small.

An improvement is made by presetting an adaptive threshold and performing an iterative location. In the improved method, the standard deviation ( $\sigma$ ) of the spectrum is calculated first, and then some strong lines are located with the threshold of  $3\sigma$ . Subsequently, the located strong lines are subtracted and the standard deviation of the spectrum is re-calculated. Some sub-strong lines are then located with the new threshold of  $3\sigma$ . The iteration will continue until the number of the located lines is up to the number of lines in Table 1. Using the improved method, ill locations can be avoided as much as possible.

### 2.2.2 Initial Identification

In this step, the first three lines are selected from both the sequence of the located lines and the theoretical lines. According to Equations (1) and (2), three parameters are defined as follows:

$$\theta = \left| \frac{\lambda_3 - \lambda_2}{\lambda_2 - \lambda_1} - \frac{x_3 - x_2}{x_2 - x_1} \right| \quad (3)$$

$$\alpha_1 = \left| \frac{V_t^3}{V_t^2} - \frac{V_o^3}{V_o^2} \right| \quad (4)$$

$$\alpha_2 = \left| \frac{V_t^1}{V_t^2} - \frac{V_o^1}{V_o^2} \right| \quad (5)$$

The three parameters are calculated and compared with the preset thresholds ( $\theta_{th}$  and  $\alpha_{th}$ ). If  $\theta \leq \theta_{th}$ ,  $\alpha_1 \leq \alpha_{th}$ ,  $\alpha_2 \leq \alpha_{th}$ , the pixel positions and wavelengths of the three

lines are recorded as initial identification. Otherwise, the next line in the sequence is added and arbitrary three lines are selected from the four lines to re-calculate the parameters. Hence there are  $C_4^3 \cdot C_4^3$  combinations of the lines selection, and the combination with minimum parameter  $\theta$  is recorded.

If the parameters of the recorded combination match the condition  $\theta \leq \theta_{th}$ ,  $\alpha_1 \leq \alpha_{th}$ ,  $\alpha_2 \leq \alpha_{th}$ , the corresponding pixel positions and wavelengths of the three lines are recorded as initial identification. Otherwise, another line is added, and the similar calculations are repeated to acquire the satisfying identification. If ten lines are calculated and the initial identification is still unsatisfying, the iteration will be interrupted. For this case, the automatic identification is thought to be unsuccessful and then a manual interactive identification has to be performed.

### 2.2.3 Iterative Identification

Utilizing the initial identification, the other lines can be identified iteratively one by one. Assuming the  $i$ th line has been identified, the similar parameters can be defined as

$$\theta' = \frac{\lambda_{i+1} - \lambda_i}{\lambda_i - \lambda_{i-1}} - \frac{x_{i+1} - x_i}{x_i - x_{i-1}} \quad (6)$$

$$\alpha' = \left| \frac{V_t^{i+1}}{V_t^i} - \frac{V_o^{i+1}}{V_o^i} \right| \quad (7)$$

Using the same thresholds ( $\theta_{th}$  and  $\alpha_{th}$ ) as in Section 2.2.2, the iterative identification is performed as follows:

1. Calculate the parameters  $\theta'$  and  $\alpha'$ .
2. If  $\theta' < -\theta_{th}$ ,  $\lambda_{i+1}$  and  $V_t^{i+1}$  are replaced by the wavelength and intensity of the next line in the theoretical lines list and then go back to step 1 to recalculate the parameters. This case means that a noise is mistaken as an emission line.
3. If  $\theta' > \theta_{th}$ ,  $x_{i+1}$  and  $V_o^{i+1}$  are replaced by the pixel position and intensity of the next line in the located lines list and then go to step 1 to recalculate the parameters. This case means that a weak emission line is missed.
4. If  $|\theta'| \leq \theta_{th}$ ,  $x_{i+1}$  and  $\lambda_{i+1}$  are considered as a possible identification. And then it is verified by the relative intensity. If  $\alpha' \leq \alpha_{th}$ ,  $x_{i+1}$  and  $\lambda_{i+1}$  are recorded in the sequence of identified lines. Otherwise, all of the variables with subscript  $i + 1$  in Equations (6) and (7) are replaced by the next corresponding values in the located lines sequence and the theoretical lines list. Afterwards, the identification restarts from step 1.

The iterative identification continues until either the last located line or the last theoretical line is identified. The recorded pixel positions and wavelengths of identified lines will be used for subsequent processing.

### 2.2.4 Evaluation and Rejection

In order to ensure the reliability, the identification is evaluated and ill identified lines are rejected. The relationship of the recorded pixel positions and wavelengths

is fitted with a five term Legendre polynomial (the selection of polynomial term will be discussed in Section 3.1). Using the fitting result, the wavelengths corresponding to the pixel positions of the identified lines are calculated. The calculated wavelengths are compared with the theoretical values to get the absolute errors. If the maximum absolute error is greater than a preset threshold ( $error_{th}$ ), the corresponding line will be rejected. The evaluation and rejection continues iteratively until the absolute errors all less than the preset threshold.

After evaluation and rejection, the automatic identification is finished. The pixel positions and wavelengths of survival lines are recorded for ultimate wavelength calibration. If the number of the surviving lines is less than a preset threshold ( $N_{th}$ ) or there is no line in a considerable wide wavelength range, the automatic identification is considered unreliable. For this case, a manual interactive identification has to be performed.

### 2.3 Ultimate Wavelength Calibration

The ultimate wavelength calibration is performed referring to the proposed method (Burles & Schlegel 2007). Utilizing the result of the automatic identification, the centers of the identified lines in the spectra of other fibers are located. As the distortion of the adjacent spectra is slight, the pixel positions recorded in Section 2.2 can be considered as the centers of the identified lines in the adjacent spectra. And then the centers are corrected by a gravity method iteratively. Using this method, the centers of the identified lines of the other spectra are located one by one along the spatial direction.

As the throughputs of different fibers have discrepancy more or less, some weak lines in the fibers with low throughput may not be located exactly enough. Thus, the centers of each identified line from different spectra are fitted with a five term Legendre polynomial, and then re-calculated by the fitted polynomial.

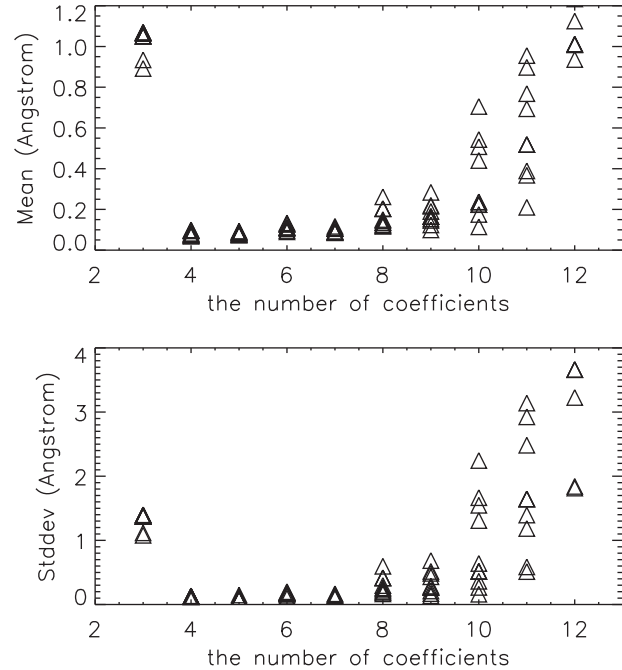
Finally, the central pixel positions and the corresponding wavelengths of each spectrum are fitted with a five term Legendre polynomial. Utilizing the fitting results, the wavelength of each pixel position of a certain spectrum can be calibrated by

$$\lambda_i = \sum_{j=0}^4 L_j(x_i)^j, \quad (8)$$

where  $L_j$  is the  $j$ th coefficient of the Legendre polynomial, and  $x_i$  is the  $i$ th pixel position. The coefficients of the Legendre polynomial are stored to be applied subsequently to the science exposures.

## 3 Experiments and Discussions

The algorithm described in Section 2.2 is programmed in Interactive Data Language (IDL), and implemented on the authors' computer (CPU: 2\*2G, EMS memory: 2G). The thresholds in the algorithm are preset as  $\theta_{th} = 0.1$ ,  $\alpha_{th} = 0.5$ ,  $error_{th} = 0.2 \text{ \AA}$ ,  $N_{th} = 20$ . Before performing



**Figure 3** The mean and standard deviation (Stddev) of the absolute errors. Each symbol represents a mean or standard deviation value calculated by all lines of a certain fiber with a certain polynomial term.

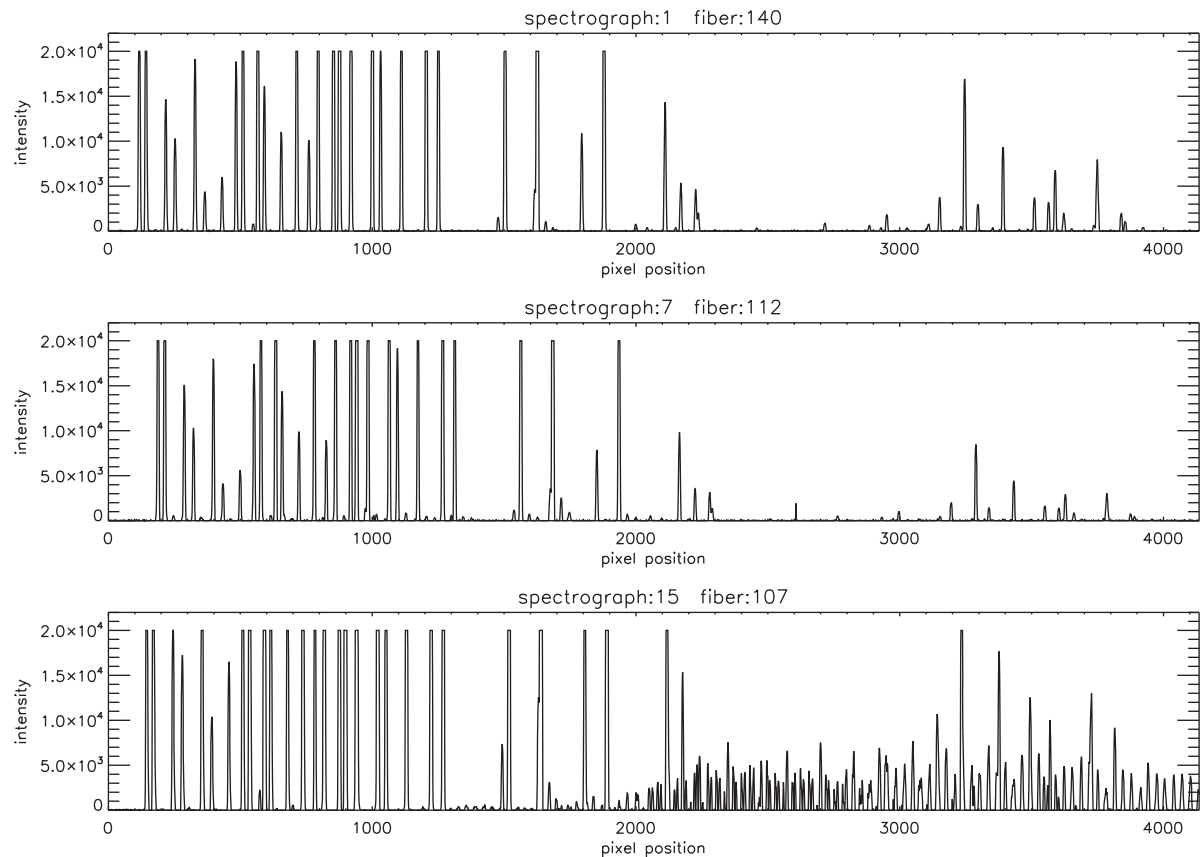
the experiments, the selection of the polynomial term is discussed (Section 3.1). Experiments are then performed on some actual spectra observed by LAMOST (Section 3.2). In order to test the robustness of the algorithm, additional experiments are implemented on some artificial spectra (Section 3.3). Finally, the ultimate wavelength calibration on the entire frame is performed and the results are compared with the alternate method.

### 3.1 Selection of Polynomial Term

Before operating the experiments, the term of the Legendre polynomial should be determined carefully to enhance the accuracy of the calibration. In this section, ten different spectra are employed to test the performance of polynomials with different terms. For each spectrum, the emission lines of the selected spectra are firstly identified as described in Sections 2.2.1 and 2.2.3. Subsequently, the relationship of the pixel position and the wavelength of the identified lines is fitted using Legendre polynomials with different terms. For each fitting, the wavelengths of the emission lines are calculated using the fitting result and then compared with their theoretical values. The mean and standard deviation of the absolute errors are calculated and plotted in Figure 3. As can be seen from Figure 3, the five term polynomial performs the best, so it is employed in the later experiments.

### 3.2 Experiments on Observed Spectra

Before starting the automatic identification, an appropriate 1D spectrum should be selected from the extracted ones. The intensity of the selected spectrum should be



**Figure 4** Three spectra used in the experiments. Some strong lines are cut off to show the weak lines clearly. The spectra have obvious differences in dispersion and intensity. Especially in the spectrum selected from the 15th spectrograph, there are a mass of false lines caused by noise.

**Table 2. Results of experiments on observed spectra**

Spectrum <sup>a</sup>	$N_L^b$	$N_I^b$	Mean	Stddev	$T$ (ms)
1–140	42	38	0.032	0.035	63
3–132	41	36	0.037	0.037	63
5–119	43	38	0.034	0.043	63
7–112	42	37	0.040	0.042	63
9–105	41	37	0.039	0.045	63
11–119	43	37	0.035	0.047	63
13–124	43	38	0.044	0.050	63
15–107	45	30	0.056	0.050	78

<sup>a</sup>Spectrum  $m-n$  means that it is extracted from fiber  $n$  of spectrograph  $m$ .

<sup>b</sup> $N_L$  is the number of the located lines, and  $N_I$  is number of the final identified lines.

considerable strong, and the position should be close to the center of the 2D frame where the distortion is slight. In this section, experiments are performed on eight spectra selected from different frames. Three of these spectra are plotted in Figure 4 to show the discrepancy in dispersion and intensity.

The selected spectra are reduced following the steps described in Section 2.2. In the step of evaluation and rejection, the mean and standard deviation of the calibrated lines are calculated. The total runtime for each spectrum is also recorded respectively. According to the experimental

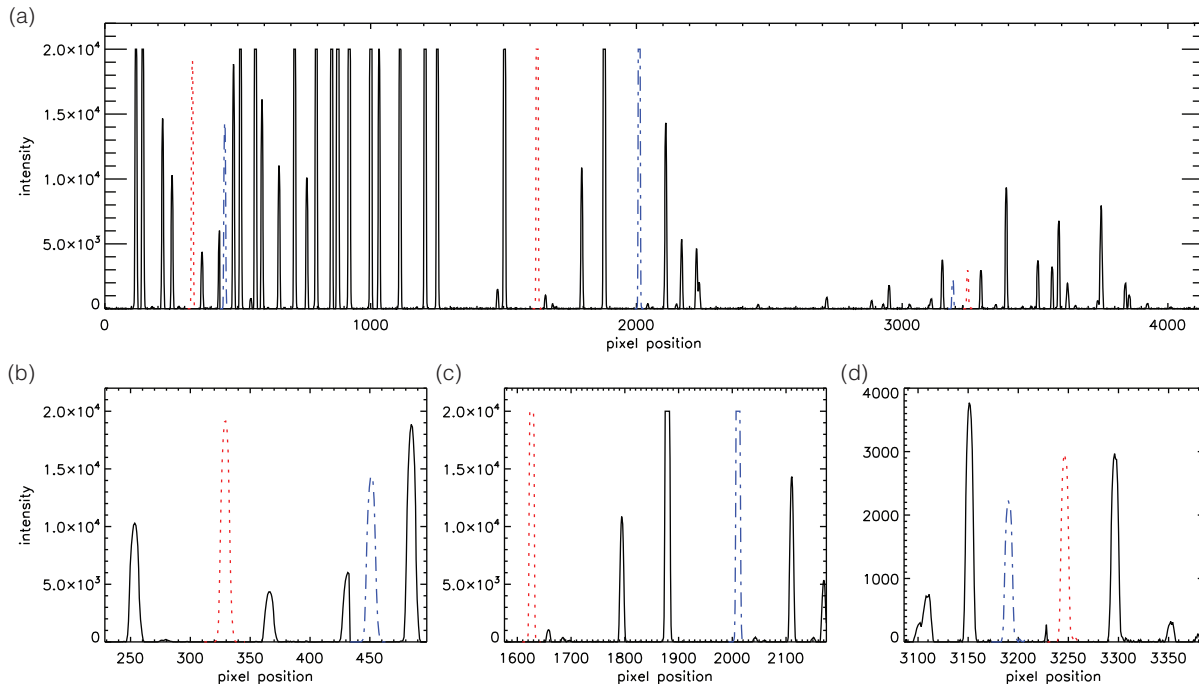
**Table 3. Pixel positions of the subtracted lines and added lines in the artificial spectra**

Spectrum	Subtracted lines	Added lines
1	217	1570
2	879	3455
3	3589	290
4	117, 1626	1950, 3210
5	760, 3391	190, 2450
6	365, 1794	75, 3675
7	253, 920, 3511	391, 1321, 3021
8	329, 1504, 3296	181, 1367, 3430
9	329, 1626, 3246	451, 2010, 3190

results listed in Table 2, the presented algorithm works well on all of the selected spectra. For each spectrum, sufficient emission lines are identified and the accuracy of the calibration is satisfying. The runtime ( $T$ ) of the algorithm for one certain spectrum is less than 0.1 s, which is much more suitable than the manual identification for the massive data reduction for LAMOST.

### 3.3 Experiments on Artificial Spectra

In order to test the robustness of the algorithm, additional experiments are implemented on some artificial spectra. In these spectra, some actual emission lines are



**Figure 5** One certain artificial spectrum used in the experiment. Subfigure (a) shows the entire spectrum, while subfigures (b–d) show parts of the spectrum in some small ranges with artificial lines. The red dotted lines present the subtracted emission lines and the blue dash-dotted lines present the added ones.

**Table 4. Results of experiments on artificial spectra**

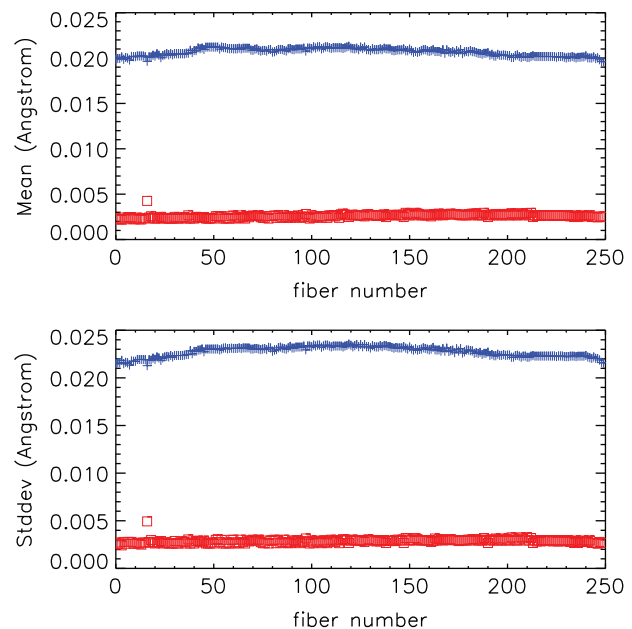
Spectrum	$N_L$	$N_I$	Mean	Stddev	$T$ (ms)
1	42	37	0.031	0.036	78
2	42	37	0.031	0.035	78
3	42	37	0.041	0.040	63
4	43	36	0.032	0.035	78
5	42	36	0.033	0.035	78
6	42	37	0.031	0.036	63
7	42	36	0.032	0.036	78
8	42	35	0.031	0.035	62
9	43	34	0.032	0.037	78

subtracted while some false ones are added. These spectra are employed to simulate the case that the spectra are influenced by strong noises. The pixel positions of the subtracted and added lines are listed in Table 3. One of the artificial spectra is plotted in Figure 5.

The artificial spectra are reduced just like Section 3.1, without any adjustment in procedures. Similar results of the experiments are listed in Table 4. Comparing with the results in Section 3.1, the identified lines are also sufficient enough, and the accuracy of the calibration almost has no difference. The runtime is still less than 0.1 second, which is much better than manual identification. Therefore, the algorithm is reliable for the data reduction of LAMOST.

### 3.4 The Ultimate Wavelength Calibration

In this section, the ultimate wavelength calibration is performed on some actual frames following the steps described in Section 2. The frames are also reduced by the



**Figure 6** The Mean and Stddev of the two methods on a certain frame. The red square symbol represents the results of the automatic method, and the blue 'plus' symbol represents the results of the alternate method. Each symbol represents a Mean or Stddev value calculated by all lines of a certain fiber.

alternate method with a preset initial wavelength solution. The Mean and Stddev of the two methods on a certain frame are plotted in Figure 6. Some other frames from different spectrographs are also reduced and the similar results are listed in Table 5.

**Table 5. The mean and Stddev of the two methods on some frames from a different spectrograph**

$S_{\text{num}}$	$\text{Mean}_A$	$\text{Stddev}_A$	$\text{Mean}_I$	$\text{Stddev}_I$
1	0.0030	0.0033	0.0092	0.0120
3	0.0036	0.0040	0.1088	0.1206
5	0.0026	0.0028	0.0206	0.0225
7	0.0032	0.0037	0.0024	0.0028
9	0.0037	0.0039	0.0649	0.0728
11	0.0031	0.0036	0.0197	0.0220
13	0.0017	0.0022	0.0031	0.0033
15	0.0044	0.0051	0.0603	0.0673

$\text{Mean}_A$  and  $\text{Stddev}_A$  are the results of the automatic method, while  $\text{Mean}_I$  and  $\text{Stddev}_I$  are the relative results of the alternate method.

Figure 6 and Table 5 show that the proposed method works stably for frames from different spectrographs, and the accuracy of the calibration is satisfying. However, the accuracy of the alternate method with the same initial wavelength solution varies too much for different frames, for some frames, the errors are too much to be accepted. Therefore, the automatic method is more applicable for the wavelength calibration of LAMOST.

#### 4 Conclusion

In this study, an adaptive algorithm is developed for the wavelength calibration of LAMOST. The new algorithm

imitates the manual interactive identification to identify the emission lines of the lamp spectra. Some experiments are performed both on the actual spectra and the artificial ones. The results of experiments show that the automatic identification is successful and the accuracy of calibration is satisfying. The runtime of the algorithm is far less than manual identification, which is much more suitable for the massive data reduction. The algorithm is also proved to be reliable for different spectra. Thus, the proposed method is applicable to the wavelength calibration of LAMOST, and it is also expected to be helpful for the data reduction of other similar instruments.

#### Acknowledgments

The authors acknowledge support from the Large Sky Area Multi-Object Fiber Spectroscopic Telescope project team. This work was funded by National Science Foundation of China (NSFC grant 10973021).

#### References

- Bai, L., Liao, N. F., Li, Z. J. & Yang, W. P., 2004, *ChOpL*, 2, 174
- Burles, S. & Schlegel, D., 2007, *SDSS Data Reduction Software*, v5.3.12, <http://das.sdss.org/software>
- Cui, X. Q., 2009, *AAS Meeting*, 213, 226.01
- Lo, E. & Fountain, A. W., 2006, *SPIE*, 6233, 62330
- Qin, H. Q., Ye, Z. F. & Luo, A. L., 2009, *PASP*, 121, 408
- Sanchez, S. F., 2006, *AN*, 327, 850
- Thompson, R. I., Bechtold, J., Black, J. H. & Martins, C. J. A. P., 2009, *NA*, 14, 379
- van Geffen, J. H. G. M. & van Oss, R. F., 2003, *ApOpt*, 42, 2739